On the Nonconcavity of Throughput in Certain Closed Queueing Networks *

Kathryn E. Stecke
Graduate School of Business Administration, The University of Michigan, Ann Arbor, MI 48109-1234, U.S.A.

Received 12 June 1984
Revised 15 March 1985 and 5 December 1985

Analytic queueing network models are being used to analyze various optimization problems such as server allocation, design and capacity issues, optimal routing, and workload allocation. The mathematical properties of the relevant performance measures, such as throughput, are important for optimization purposes and for insight into system performance.

We show that for closed queueing networks of m arbitrarily connected single server queues with n customers, throughput, as a function of a scaled, constrained workload, is not concave. In fact, the function appears to be strictly quasi-concave. There is a constraint on the total workload that must be allocated among the servers in the network. However, for closed networks of two single server queues, we prove that our scaled throughput is concave when there are two customers in the network and strictly quasi-concave when there are more than two customers. The mathematical properties of both the scaled throughput and reciprocal throughput are demonstrated graphically for closed networks of two and three single server queues.

Keywords: Closed Queueing Networks, Flexible Manufacturing Systems, Performance Evaluation, Nonconcavity of Throughput.

Kathryn E. Stecke was born in Boston, MA, in June 1950. She received the B.S. degree in Mathematics from Boston State College in 1972. She received an M.S. in Applied Mathematics in 1974 and an M.S. in Industrial Engineering from Purdue University in 1977. She is now an Associate Professor of Operations Management at the Graduate School of Business Administration at The University of Michigan. She is the Editor of the *International Journal of Flexible Manufacturing Systems*. She has authored numerous papers on various aspects of the planning and scheduling of flexible manufacturing systems in numerous journals including *The FMS Magazine, Material Flow, International Journal of Production Research, European Journal of Operational Research, Annals of Operations Research, Management Science, Operations Research* and several proceedings and book contributions. She is Chairperson of the First (and Second) ORSA/TIMS Conference on "Flexible Manufacturing Systems: Operations Research Models and Applications", held in Ann Arbor, Michigan in August 1984 (and 1986). She is currently on leave at General Motors Research Laboratories and spent Fall 1984 at the Centre d'Études et de Recherches de Toulouse.

* This research was supported in part by NSF Grant No. ECS 8406407 as well as by a Grant from the Graduate School of Business Administration of The University of Michigan.

North-Holland
Performance Evaluation 6 (1986) 293–305

1. Introduction

Closed queueing network models have recently been used to analyze design issues and planning problems of both computer systems and flexible manufacturing systems. Throughput, a main performance measure of interest, can be defined as a complex, nonlinear function of several system parameters. The mathematical and qualitative properties of this function are of interest for optimization and performance evaluation purposes. For example, in the problem of maximizing throughput subject to a set of contraints, it is necessary to know if a local maximum is a global maximum. In addition, studying the qualitative properties of throughput is also useful for the analytic insight that is provided.

Various versions of this problem have been reported in the computer science literature. For example, Trivedi and Kinicki [25], Trivedi and Wagner [27], Trivedi, Wagner, and Sigmon [28], Trivedi and Sigmon [26], and Kobayashi and Gerla [8] maximize throughput in central server, single class, closed queueing networks (CQN) with a single server at each node subject to various budgetary limitations (cost constraints). The various studies optimize different parameters (decision variables) such as service rate (of a CPU, say), capacity of servers (I/O devices), device speeds, routing, and main memory size, often subject to a budget constraint. These parameters relate cost considerations to performance. All of these studies prove the convexity of reciprocal throughput in order to insure that the maximum

0166-5316/86/$3.50 © 1986, Elsevier Science Publishers B.V. (North-Holland)

throughput (minimum average delay or response time) is global, and not just a local optimum. To prove convexity, these studies use the results of Price [12], who proved convexity for a particular scaled version of reciprocal throughput.

However, the reciprocal of a convex function is not necessarily concave [10]. In fact, the reciprocal of a convex function can be either quasi-concave or quasi-convex. There could be some benefits and additional insights from investigating the mathematical properties of throughput directly. For example, Suri [24] analyzes the sensitivity (and bounds on sensitivity) of throughput to variations in workoad, as well as other properties of throughput.

There have not been many studies that analyze throughput directly. Kenevan and Von Mayrhauser [7] show that throughput is a log convex function of the number of items in a closed, single class, network of an arbitrary number of single and instant servers. They also prove that reciprocal throughput is a convex function of the relative utilization of the servers. This is a generalization of Price's [12] proof.

The following studies provide results concerning optimal solutions (workload allocations and server configurations) to problems of maximizing throughput in both single server and multiserver CQNs.

Kobayashi and Gerla [8] determine the optimal routing to maximize throughput in central-server, single server, single class CQNs. Stecke and Morin [21] and Yao [29] show that balancing workloads, for various scalings, maximizes throughput in single server, arbitrarily-connected CQNs. Shanthikumar and Stecke [15] prove that balancing the workloads in single server CQNs minimizes in-process inventory under various strategies to release items to the network.

For multiserver CQNs, Stecke and Solberg [23] and Yao [29] prove that balancing workloads per queue maximizes throughput when each queue has the same number of servers. However, Stecke and Solberg [23] also show that when the number of servers in each queue is not the same in multiserver CQNs, the throughput is maximized by a unique unbalanced workload per server. In fact in this situation, the throughput function appears not only to be not concave, but not symmetric as well. Unbalanced optimal workload allocation ratios can be found at which the workload per server should be maintained to maximize throughput for these networks of unbalanced multiserver queues. These allocation ratios can serve as input to more detailed workload allocation problems that are solved using more detailed (mathematical programming) models (see, for example, Berrada and Stecke [3] and Stecke [18,19,20]).

We consider here a particular product form, noncentral server CQN of arbitrarily-connected single server queues, of which the central server model is a special case. Rather than the budgetary constraints of the previous studies, we impose a constraint on the total workload in the system. The motivation for our particular CQN model is provided in the studies of optimal workload allocation and server (machine) allocation in flexible manufacturing systems (FMSs). In particular, we show, contrary to previous conjectures [14,17], that throughput (or production rate) is not concave as a function of workload.

In this paper, we show that throughput, as a function of the ratio of the 'workload' (service demand) at a server to the sum of workloads is quasi-concave and not concave. Since Price [12] and Kenevan and Von Mayrhauser [7] do *not* consider throughput to be a function of the same quantity (a ratio of server to total workload) and do *not* constrain the total workload to be allocated, their results do not necessarily apply. However, there is evidence that the reciprocal throughput function studied here *is* convex, despite the particular scaling of workload and throughput.

The plan of the paper is the following. In Section 2, the CQN model is defined. We prove the nonconcavity results by induction in Section 3. First, the concavity of throughput is proven for a closed network of two single server queues with two customers. Then, the nonconcavity is established numerically for a network of two single server stations with n (greater than two) customers. Next, strict quasi-concavity of throughput is established for a CQN with two single servers. A concave function is also quasi-concave; however, we also show in Section 3 that this scaled throughput function is not concave for m queues and n customers. For definitions of generalized concavity, see Bazaraa and Shetty [2] or Martos [10]. In Section 4, some evidence that our scaled version of reciprocal throughput is convex is provided. If this is true, we can prove that throughput

is strictly quasi-concave. Section 5 concludes with a brief summary.

2. The closed queueing network model

The product form CQN that is considered here consists of m arbitrarily connected single servers, of which the central server model is a special case. There are always n items being processed in the system. The average processing time of an item at station i is t_i, $i = 1, 2, \ldots, m$. The routing of items among the stations is arbitrary. The routing can be described by visit frequencies, or relative arrival rates, q_i, where q_i can be the probability that the next server visited is i. In addition, the q_i can be provided by the traffic equations, $q_i = \sum_{j=1}^{m} p_{ji} q_j$. Details of other routing possibilities can be found in [22].

The queueing discipline can be either FCFS, infinite server, LCFS preempt-resume, processor sharing (see [1]), random selection, or one developed by Kelly [6] that allows an arbitrary distribution to be defined at each server. The service time distribution is arbitrary, except for FCFS servers, which require exponential service times.

The usual measure of relative workload assigned to server i is w_i [4,13,16], which is defined as the product of visit frequency and average processing time, or $w_i = q_i t_i$. These workloads are relative since the q_i's need not sum to one.

For our purposes, w_i was scaled, where $\sum_{j=1}^{m} q_j t_j / m$ is the average workload per server, to provide:

$$X_i = q_i t_i \Big/ \left[\left(\sum_{j=1}^{m} q_j t_j \right) / m \right]. \tag{1}$$

This particular constraint on workload was chosen for many reasons associated with determining qualitative properties of optimal allocations of servers and workloads in flexible manufacturing systems (see [20,22,21,23] for details on these studies).

The state of the system is given by $\tilde{n} = (n_1, \ldots, n_m)$, where n_i is the number of items at server i, both those waiting and in process. For all i, we have $n_i \in \{0, 1, \ldots, n\}$ and $\sum_{i=1}^{m} n_i = n$. The steady-state probability of being in state $\tilde{n}$ is $p(\tilde{n}) = p(n_1, n_2, \ldots, n_m)$, which has the product form solution

$$p(\tilde{n}) = \frac{1}{G(m, n; X)} X_1^{n_1} X_2^{n_2} \cdots X_m^{n_m},$$

where

$$G(m, n; X) = \sum_{\substack{n_1 + n_2 + \cdots + n_m = n \\ n_i \geq 0}} X_1^{n_1} X_2^{n_2} \cdots X_m^{n_m},$$

$$i = 1, 2, \ldots, m. \tag{2}$$

Throughput can be defined as a function of $G(m, n; X)$, which in turn is a function of assigned workload, X_i. In fact, for a particular scaling of q_i, the throughput, or production rate, $\Pr(m, n; X)$, is given by [13]

$$\begin{aligned} \Pr(m, n; X) &= \frac{G(m, n-1; X)}{G(m, n; X)} \\ &= \left(\sum_{\substack{n_1 + n_2 + \cdots + n_m = n-1 \\ n_i \geq 0}} X_1^{n_1} X_2^{n_2} \cdots X_m^{n_m} \right) \\ &\quad \times \left(\sum_{\substack{n_1 + n_2 + \cdots + n_m = n \\ n_i \geq 0}} X_1^{n_1} X_2^{n_2} \cdots X_m^{n_m} \right)^{-1} \end{aligned} \tag{3}$$

The throughput for two single servers and any number of items is

$$\begin{aligned} \Pr(2, n; X) &= \left(\sum_{n_1 + n_2 = n-1} X_1^{n_1} X_2^{n_2} \right) \Big/ \left(\sum_{n_1 + n_2 = n} X_1^{n_1} X_2^{n_2} \right) \\ &= \left(\sum_{n_1 = 0}^{n-1} X_1^{n_1} (2 - X_1)^{n-1-n_1} \right) \\ &\quad \times \left(\sum_{n_1 = 0}^{n} X_1^{n_1} (2 - X_1)^{n-n_1} \right)^{-1}, \end{aligned}$$

since $X_1 + X_2 = m = 2$ (with our scaling)

$$= \left(X_1^n - (2 - X_1)^n \right) / \left(X_1^{n+1} - (2 - X_1)^{n+1} \right), \tag{4}$$

by dividing both numerator and denominator by $(2 - X_1) - X_1 = 2(1 - X_1)$.

Throughput, as given in equation (3), is difficult to characterize analytically. However, it can

be evaluated numerically using Buzen's efficient algorithm [4].

3. (Non)concavity of throughput

In this section, first the concavity of throughput is proven for a closed network of two single server stations with two items. Then, strict quasi-concavity, but nonconcavity, of throughput is established for a network of m ($\geqslant 3$) items).

For a network of two single servers with two items, from equation (4):

$$\Pr(2, 2; X) = \frac{X_1^2 - X_2^2}{X_1^3 - X_2^3} = \frac{X_1 + X_2}{X_1^2 + X_1 X_2 + X_2^2}.$$

Substituting $X_2 = 2 - X_1$, simplifying, and then dropping the subscript, we obtain

$$\Pr(2, 2; X) = \frac{2}{4 - 2X + X^2}.$$

Theorem 3.1. $\Pr(2, 2; X)$ *is a concave function.*

Proof. Taking the first derivative of $\Pr(2, 2; X)$ yields

$$\frac{\mathrm{d}\Pr(2, 2; X)}{\mathrm{d}X} = \frac{-2(2X-2)}{[4 - X(2-X)]^2} = \frac{-4(X-1)}{[4 - X(2-X)]^2} = \frac{-4(X-1)}{(X^2 - 2X + 4)^2}.$$

Setting $\Pr'(2, 2; X) = 0$ yields $X = 1$. Now,

$$\frac{\mathrm{d}^2\Pr(2, 2; X)}{\mathrm{d}x^2} = \frac{-4(X^2 - 2X + 4) - 4(4 - 4X)(X-1)}{(X^2 - 2X + 4)^3}$$

$$= \frac{-4X^2 + 8X - 16 + 16(X^2 - 2X + 1)}{(X^2 - 2X + 4)^3}$$

$$= \frac{4(3X^2 - 6X)}{(X^2 - 2X + 4)^3} = \frac{12X(X-2)}{(X^2 - 2X + 4)^3}.$$

Setting $\Pr''(2, 2; X) = 0$, the points of inflection are at $X = 0$ and $X = 2$. Note that, for every $X \in [0, 1)$, $\Pr'(2, 2; X) > 0$, which implies that $\Pr(2, 2; X)$ is increasing on $[0, 1)$; for every $X \in (1, 2]$, $\Pr'(2, 2; X) < 0$, which implies that $\Pr(2, 2; X)$ is decreasing on $(1, 2]$. □

Theorem 3.2. $\Pr(2, n; X)$ *is not concave for* $n \geqslant 3$.

Proof

$$\Pr(2, n; X) = \left.\frac{X_1^n - X_2^n}{X_1^{n+1} - X_2^{n+1}}\right|_{X_1 + X_2 = 2} = \frac{X_1^n - (2 - X_1)^n}{X_1^{n+1} - (2 - X_1)^{n+1}}.$$

Again, the subscript is suppressed for convenience. Taking the derivative with respect to X yields

$$\frac{\mathrm{d}\Pr(2, n; X)}{\mathrm{d}X} = \frac{\left[X^{n+1} - (2-X)^{n+1}\right] n\left[X^{n-1} + (2-X)^{n-1}\right]}{\left[X^{n+1} - (2-X)^{n+1}\right]^2},$$

$$- \frac{\left[X^n - (2-X)^n\right](n+1)\left[X^n + (2-X)^n\right]}{\left[X^{n+1} - (2-X)^{n+1}\right]^2},$$

which upon rearranging yields

$$= \frac{-X^{2n} + (2-X)^{2n} + 4nX^{n-1}(2-X)^{n-1}(X-1)}{\left[X^{n+1} - (2-X)^{n+1}\right]^2}.$$

Evaluating at $X=1$, $\Pr'(2, n; X) = 0/0$. Upon two applications of l'Hospital's rule we obtain

$$\Pr'(2, n; 1) = \frac{4n(n-1)(-2) + 4n(n-1) + 4n(n-1)}{2(n+1)^2 4} = \frac{0}{8(n+1)^2} = 0.$$

Therefore, $X=1$ is a critical point.

Taking the second derivative with respect to X and rearranging we obtain

$$\frac{d^2\Pr(2, n; X)}{dX^2} = \frac{2\left[X^{3n} - (2-X)^{3n} + X^{n-2}(2-X)^{2n-1}\left(X^3 - (8n+2)X^2 + (-4n^2+8n)X + 4n^2 - 4n\right)\right.}{\left[X^{n+1} - (2-X)^{n+1}\right]^3} + \frac{\left.X^{2n-1}(2-X)^{n-2}\left(X^3 + (8n-4)X^2 + (-4n^2-24n+4)X + 4n^2 + 20n\right)\right]}{\left[X^{n+1} - (2-X)^{n+1}\right]^3}. \quad (5)$$

The throughput function is now demonstrated graphically to be

(i) convex for $X_1 \in [0, X']$, $X' < 1$,
(ii) concave for $X_1 \in [X', X'']$, $X'' > 1$, and
(iii) convex for $X_1 \in [X'', 2]$.

Then there are three points of inflection: at X', 1, and X''. Moreover, X' and X'' are symmetric about the point $X=1$.

The points of inflection of $\Pr(2, n; X)$ can be found by setting the numerator of the second derivative of the nonlinear equation (5) equal to zero and solving for the roots. Two different IMSL [5] routines, called ZREAL1 (see [11,9]) and ZREAL2, were used to find the roots. Both were used as a check on accuracy and to help note any numerical or roundoff problems. Each routine finds N real zeros of a function $F(Y)$. The routines were set up to search for 5; each always found only three roots, including and symmetric about $X=1$.

There are two convergence criteria necessary. X_i is a root if

(i) $|F(Y_i)| \leqslant \text{EPS}$,

(ii) $\dfrac{Y_i^{m+1} - Y_i^m}{Y_i^m} < 10^{-\text{NSIG}}$.

In the program, EPS (epsilon) was set equal to 1.0 E–8 and NSIG = 5. The roots remained the same for both ESP = 1.0E–5 and 1.0E–8, which implies that sufficient accuracy was attained. Both routines found the same roots, as seen both in Table 1 and graphically in Fig. 1. The graph and points of inflection show that $\Pr(2, n; X)$ is not concave on $[0, 2]$. □

However, both Table 1 and Fig. 1 indicate that throughput is *strictly quasi-concave* with a global maximum at X = 1.

Table 1
Points of inflection and approximation for $n = 2, 3, 4, 5, 10,$ and 99

	ZREAL1	ZREAL2	$(2n-3)/(2n+1)$
$n=2$	$X = 0$, 1, and 2	$X = 0$, 1, and 2	
$n=3$	0.42265	0.42265	$\frac{3}{7} = 0.4286$
	1.0	1.0	
	1.57735	1.57735	
$n=4$	0.55452	0.55452	$\frac{5}{9} = 0.5555$
	1.0	1.0	
	1.44548	1.44548	
$n=5$	0.62943	0.62943	$\frac{7}{11} = 0.6366$
	1.0	1.0	
	1.37057	1.37057	
$n=7$	0.71563	0.71563	$\frac{11}{15} = 0.7333$
	1.0	1.0	
	1.28437	1.28437	
$n=10$	0.78377	0.78377	$\frac{17}{21} = 0.8095$
	1.0	1.0	
	1.21623	1.21623	
$n=25$	0.89339	0.89339	$\frac{47}{51} = 0.9215$
	1.0	1.0	
	1.10661	1.10661	
$n=99$	0.9649	0.964903	$\frac{195}{199} = 0.9799$
	1.0	1.0	
	1.0351	1.035097	

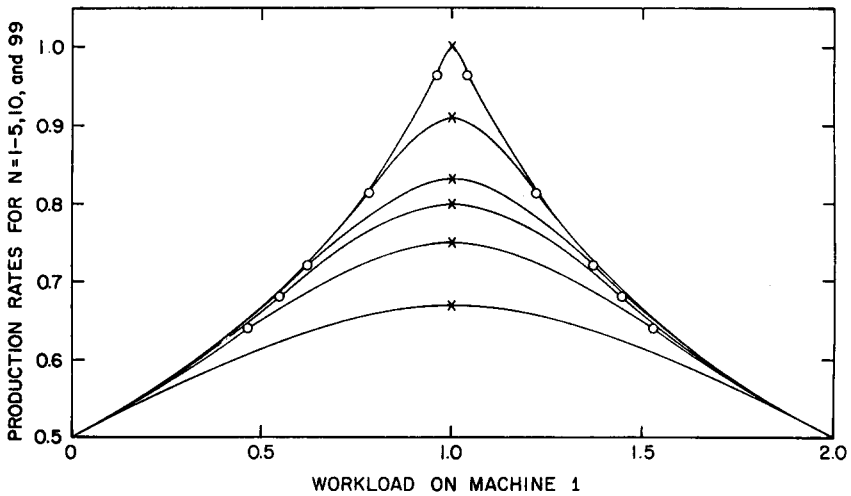

Fig. 1. Graph of $\Pr(2, n; X)$, $n = 2, 3, 4, 5, 10$, and 99: maxima ($\times$) and points of inflection ($\circ$).

Theorem 3.3. *For $n > 2$, $\Pr(2, n; X)$ is a strictly quasi-concave function on the interval $X_i \in [0, 2]$, $i = 1, 2$.*

Proof. For each $n > 2$, there are three critical points, one at $X_i = 1$ that gives the maximum $\Pr(2, n; X)$, and the remaining two symmetric

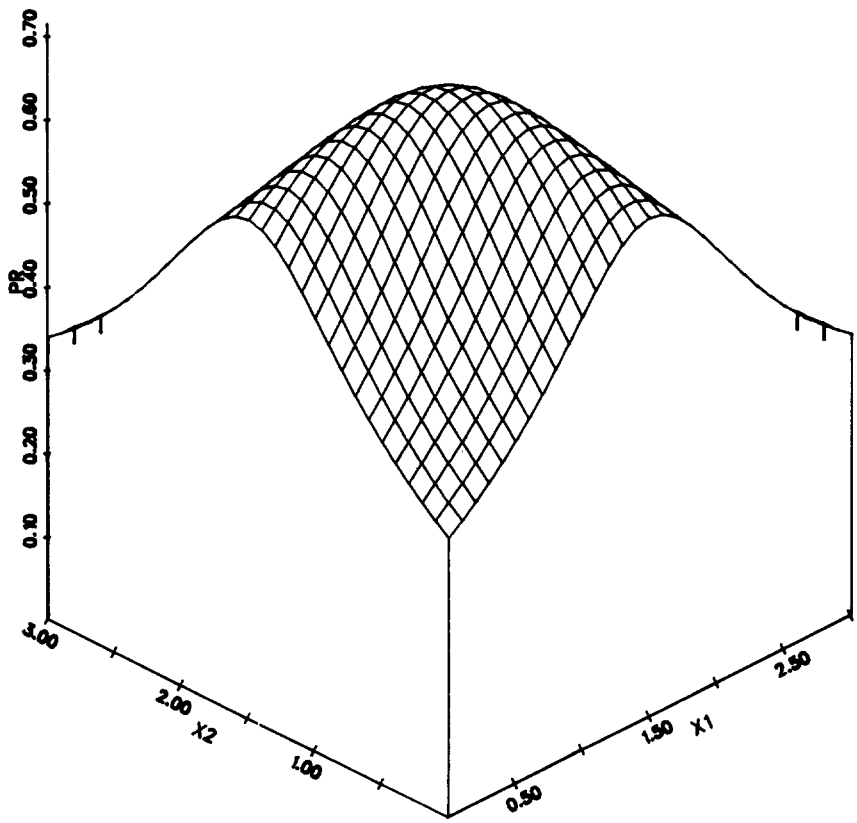

Fig. 2. Graph of $\Pr(3, 5; X)$, $X_{1,2} \in [0, 3]$.

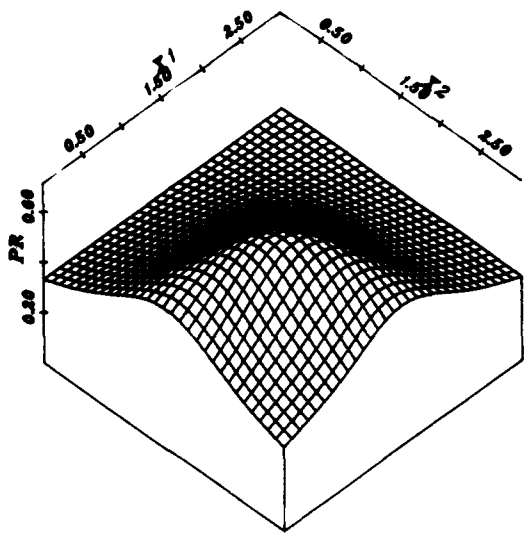

Fig. 3. Graph of $\Pr(3,5;X)$, $X_{1,2} \in [0,3]$.

about the line $X_i = 1$. The function is increasing for $X_1 \in [0,1)$ and decreasing for $X_1 \in (1,2]$. Therefore, throughput is quasi-concave for $X_1 \in [0, 2]$ with the unique global maximum at $X_1 = X_2 = 1$. □

The last column of Table 1, labeled $(2n-3)/(2n+1)$, shows the results of the attempt to provide a simple function that would closely approximate the values of the roots of $\Pr(2, n; X)$, or the points of inflection.

We now show that throughput is *not concave*.

Theorem 3.4. $\Pr(m, n; X)$ *is not concave for any* $m \geqslant 2$ *and* $n \geqslant 3$.

Proof. Consider the throughput function for any m or n:

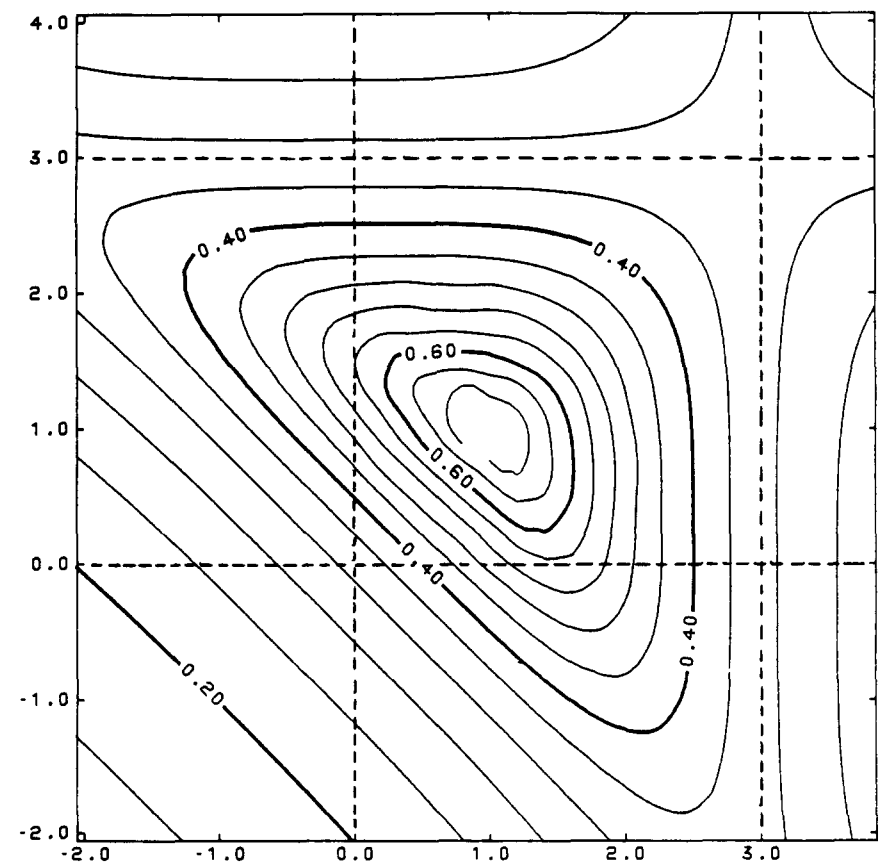

Fig. 4. Contour map of $\Pr(3,5;X)$.

$$\Pr(m, n; X) = \left(\sum_{n_1+n_2+\cdots+n_m=n-1} X_1^{n_1} X_2^{n_2} \cdots X_m^{n_m} \right) \times \left(\sum_{n_1+n_2+\cdots+n_m=n} X_1^{n_1} X_2^{n_2} \cdots X_m^{n_m} \right)^{-1}.$$

Evaluating $\Pr(m, n; X)$ along any hyperplane such that $X_i = 0$ for $m-2$ of the i, say, for $i = 1, 2, \ldots, m-2$, we have

$$\Pr(m, n; X) = \left(\sum_{n_{m-1}+n_m=n-1} X_{m-1}^{n_{m-1}} X_m^{n_m} \right) \times \left(\sum_{n_{m-1}+n_m=n} X_{m-1}^{n_{m-1}} X_m^{n_m} \right)^{-1}$$

$$= \left(\sum_{n_{m-1}=0}^{n-1} X_{m-1}^{n_{m-1}} (m - X_{m-1})^{n-1-n_{m-1}} \right) \times \left(\sum_{n_{m-1}=0}^{n} X_{m-1}^{n_{m-1}} (m - X_{m-1})^{n-n_{m-1}} \right)^{-1}$$

$$= \frac{X_{m-1}^n - (m - X_{m-1})^n}{X_{m-1}^{n+1} - (m - X_{m-1})^{n+1}}$$

$$= \frac{X_{m-1}^n - X_m^n}{X_{m-1}^{n+1} - X_m^{n+1}},$$

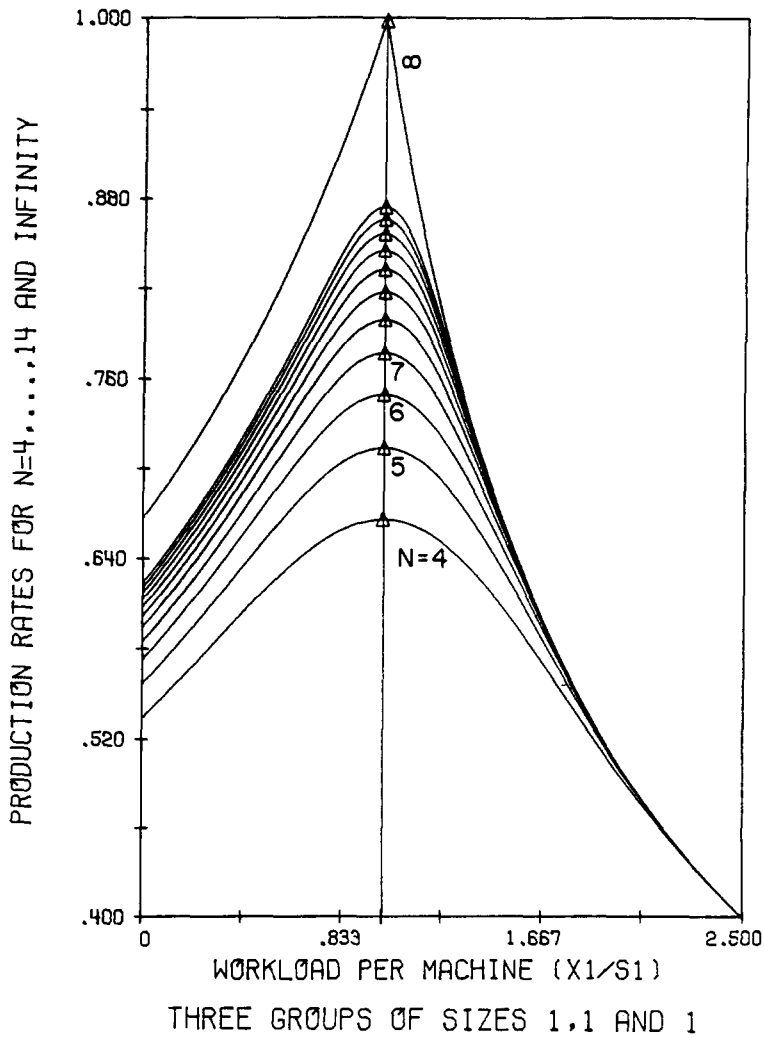

Fig. 5. A one-dimensional slice of $\Pr(3, n; X)$ as a function of server 1 for $n = 4, 5, \ldots, 14$ and ∞.

since $X_1 + X_2 + \cdots + X_m = X_{m-1} + X_m = m$.

But this is the same form as $\Pr(2, n; X)$, which has already been shown to be not concave in Theorem 3.2. □

Figs. 2 to 7 help to further demonstrate and clarify the behaviour of throughput. Figs. 2 and 3 are different views of a three-dimensional graph of $\Pr(3,5; X)$. (Numerous other plots of $\Pr(3, n; X)$ for many values of n are very similar in form to these.) Both figures show the function over its entire range of relevant workload values: since there are three singles server queues in the closed network, our scaling ensures that $X_1 + X_2 + X_3 = 3$, with each $X_i \in [0,3]$. The quasi-concavity can be seen as the function dips near the extreme boundary points.

Fig. 4 appears to demonstrate some bizarre behaviour of the throughput function, particularly outside the dashed box. The function appears to change direction. However, the function is well behaved within the dashed lines, which define the relevant range for our scaled workload.

Fig. 5 interestingly demonstrates the nonsymmetry of a one-dimensional slice of $\Pr(3, n; X)$ over a range of n, despite the symmetry of the entire function. Fig. 5 also shows the strict quasi-concavity of throughput as a function of workload.

Figs. 6 and 7 also show the strict quasi-concavity of the production function. When contrasted with Figs. 2 and 3, the behaviour is seen to exaggerate as n increases. In this example, n doubled, from five to ten customers. The closed network is more congested.

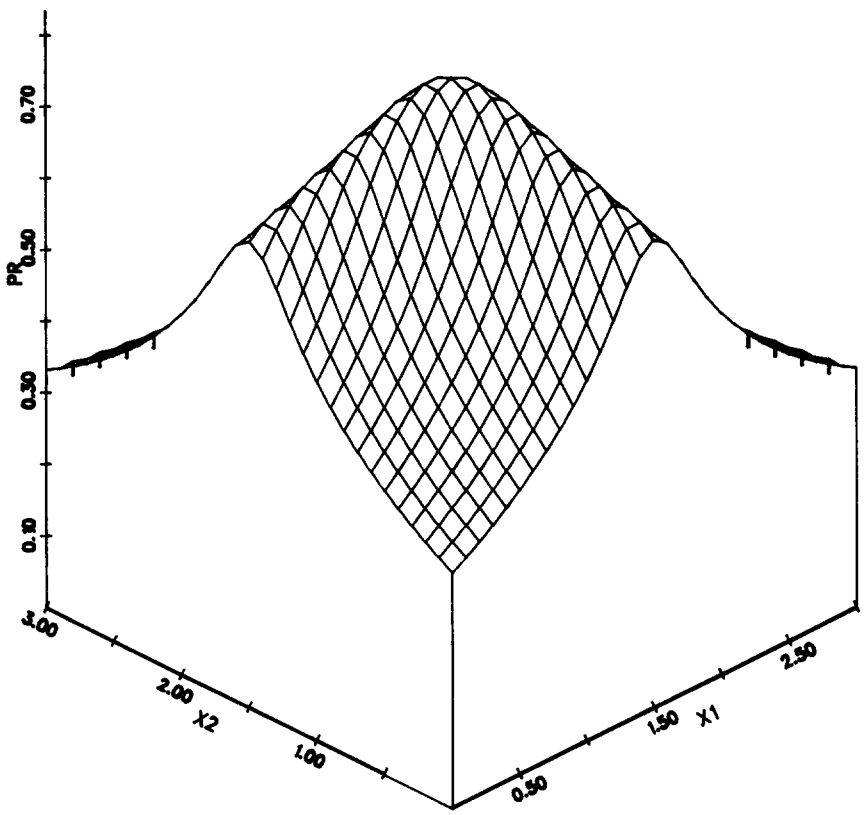

Fig. 6. Graph of $\Pr(3,10; X)$, $X_{1,2} \in [0,3]$.

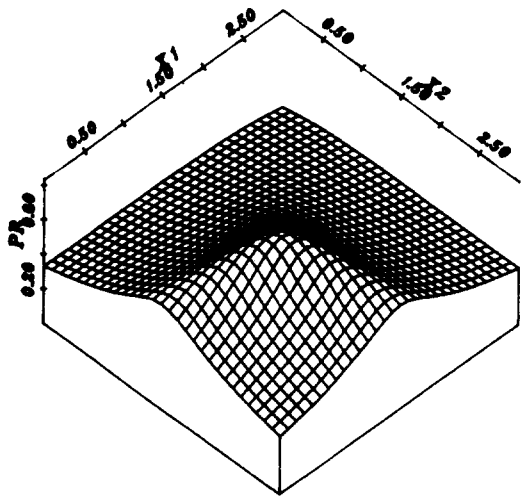

Fig. 7. Graph of $\Pr(3,10;X)$, $X_{1,2} \in [0,3]$.

4. Reciprocal throughput

For certain single server queueing networks, reciprocal throughput has been shown to be convex (see, for example, [12]). However, Price [12] does not consider throughput to be a function of the same quantities that are considered in this paper and does not consider the same total workload constraint. Hence, his results do not necessarily apply to our scaled versions of workload and reciprocal throughput.

However, although we have not formally proven convexity, we can offer some computational evidence that our scaled reciprocal throughput function is also convex. In particular, Fig. 8 demonstrates convexity for a closed network of two

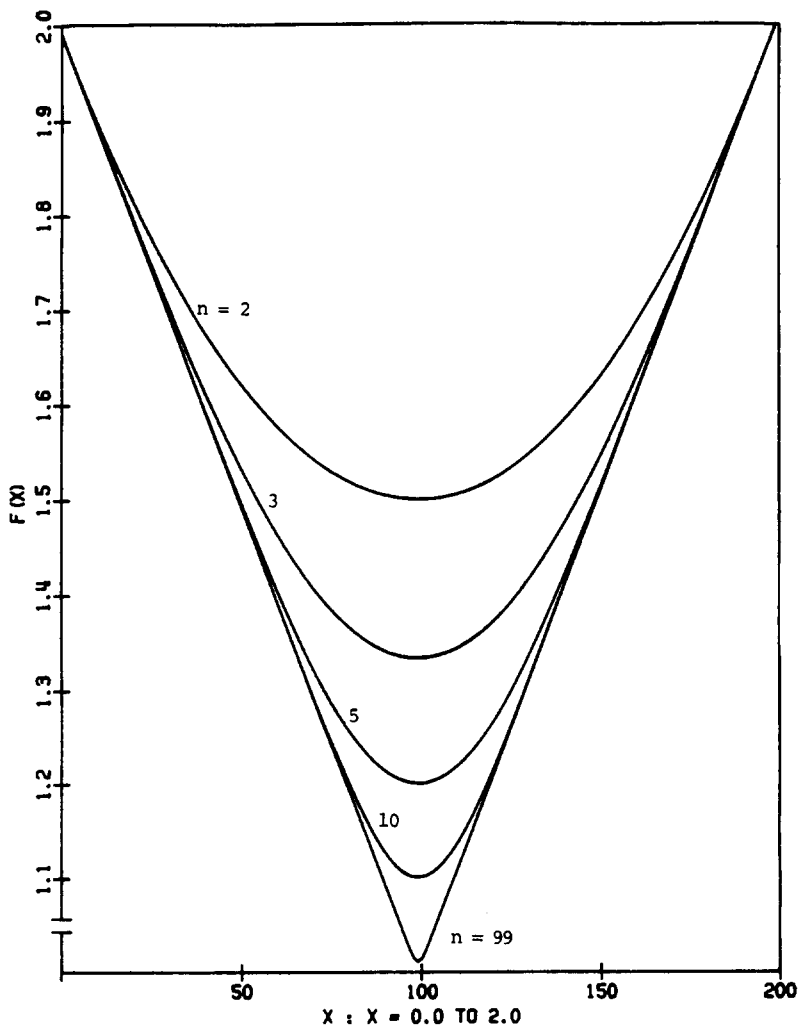

Fig. 8. Graph of $\Pr(2,n;X)^{-1}$, for a variety of n.

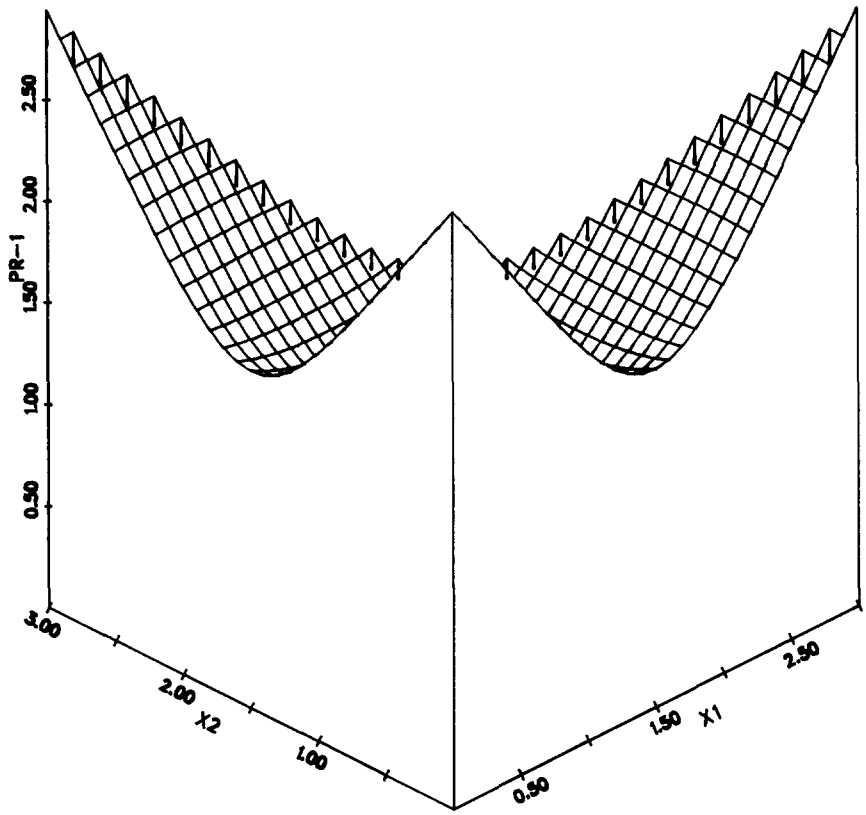

Fig. 9. A three-dimensional graph of $\Pr(3,5;X)^{-1}$, for $X_i \in [0,3]$, $i=1, 2$.

single server queues, for a variety of n, ranging from $n=2$ up to 99.

Also, Fig. 9 demonstrates the convexity for a closed network of three single server queues with the number of customers, n, equal to 5. Graphs for other values of n are similar. Finally, Fig. 10 provides some values, via a contour graph, for the reciprocal throughput function, $\Pr(3,5;X)^{-1}$, over the relevant range of scaled workload, $X_i \in [0,3]$, for $i=1, 2, 3$.

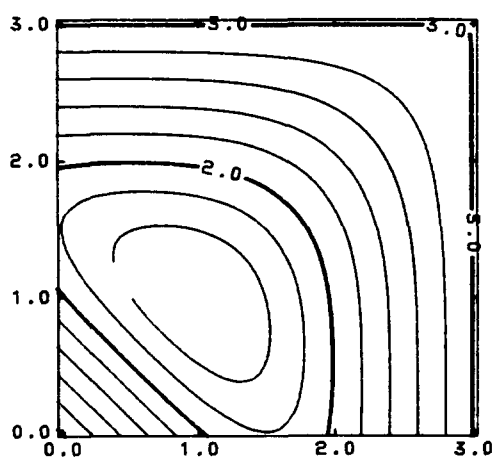

Fig. 10. Contour map of $\Pr(3,5;X)^{-1}$, for $X_i \in [0,3]$, $i=1, 2$.

These figures and many other similar graphs provide evidence that reciprocal throughput is convex. If this observation is true, we can use some previous results in mathematical programming to prove directly that throughput is strictly quasi-concave.

In particular, the following result that is stated as Theorem 4.1 can be proven in several ways. One proof can use [2, 3.39, p. 116]. Another can use [10, Variant H of Table 3.4, p. 63]. Our proof will use the latter.

Theorem 4.1. *Throughput is strictly quasi-concave, if reciprocal throughput is convex.*

Proof. Variant H of Martos [10] can be stated as follows: A function $f(x)/g(x)$ is strictly quasi-

concave if $f(x)$ is concave and nonnegative and $g(x)$ is convex and positive.

Let $f(x) = 1$ and $g(x)$ be reciprocal throughput. Then $f(x)$ is clearly concave and nonnegative for all x, and $g(x)$ is convex by assumption and positive.

Hence throughput is strictly quasi-concave. □

We note that without our particular scaling of workload (expressed via the constraint: $X_1 + X_2 + \cdots + X_m = m$), our reciprocal throughput function given by equation (3) $(\Pr(m, n; X)^{-1})$ can be shown to be convex. One proof would mimic that found in [8].

5. Summary

We have attempted to provide some mathematical and qualitative insights into a particularly useful scaled version of both throughput itself and reciprocal throughput as functions of a particular scaled workload measure. As a result, throughput is also scaled. In fact, it represents a probability: all values lie between zero and one (see [22]). To our knowledge, such properties concerning the generalized concavity, of throughput in particular, have not previously been investigated. This is, in part, because reciprocal throughput has been easier to get a handle on and is also better behaved.

If this particular, scaled, reciprocal throughput function is formally proven to be convex, Theorem 4.1 is required to characterize throughput itself directly, since the reciprocal of a convex function can be either quasi-concave or quasi-convex (recall [10, Table 3.4, p. 63]). Then, throughput is strictly quasi-concave.

Of course, if additional, general information can be discovered about either function, some of that information can be useful, for example, for performance evaluation or optimization purposes or general insights into the behaviour of the functions.

References

[1] F. Baskett, K.M. Chandy, R.R. Muntz and F.G. Palacios, Open, closed, and mixed networks of queues with different classes of customers, J. Assoc. Comput. Mach. 22 (2) (1975) 248–260.

[2] M.S. Bazaraa and C.M. Shetty, Nonlinear Programming: Theory and Algorithms (Wiley, New York, 1979).

[3] M. Berrada and K.E. Stecke, A branch and bound approach for machine load balancing in flexible manufacturing systems, Management Sci. 30 (10) (1986) 1316–1335.

[4] J.P. Buzen, Computational algorithms for closed queueing networks with exponential servers, Comm. Assoc. Comput. Mach. 16 (9) (1973) 527–531.

[5] IMSL (International Mathematical and Statistical Library) Reference Manual, IMSL, Inc., Houston, TX, 1979.

[6] F.P. Kelly, Reversibility and Stochastic Networks (Wiley, New York, 1979).

[7] J.R. Kenevan and A.K. von Mayrhauser, Convexity and concavity properties of analytic queueing models for computer systems, in: E. Gelenbe, ed., Performance '84 (Elsevier Science Publishers B.V.—North-Holland, Amsterdam, 1984) 361–375.

[8] H. Kobayashi and M. Gerla, Optimal routing in closed queueing networks, ACM Trans. Comput. Systems 1 (4) (1983) 294–310.

[9] B. Leavenworth, Algorithms 25: Real zeros of an arbitrary function, Comm. Assoc. Comput. Mach. 13 (1960).

[10] B. Martos, Nonlinear Programming: Theory and Methods (North-Holland, Amsterdam, 1975).

[11] D.E. Muller, A method for solving algebraic equations using an automatic computer, Math. Tables and Aids to Computation 10 (1956) 208–215.

[12] T.G. Price, Probability models of multiprogrammed computer systems, Ph.D. Dissertation, Dept. of Electrical Engineering, Stanford Univ., Stanford, CA, December 1974.

[13] M. Reiser and H. Kobayashi, Horner's role for the evaluation of general closed queueing networks, Comm. Assoc. Comput. Mach. 18 (10) (1975) 592–593.

[14] G. Secco-Suardo, Optimization of a closed network of queues, Rept. No. ESL-FR-834-3, Electronic Systems Laboratory, M.I.T., Cambridge, MA, 1978.

[15] J.G. Shanthikumar and K.E. Stecke, Reducing work-in-process inventory in certain classes of flexible manufacturing systems, Europ. J. Oper. Res. 26 (2) (1986) 266–271.

[16] J.J. Solberg, A mathematical model of computerized manufacturing systems, Proc. Internat. Conf. on Production Research, Tokyo, Japan, 1977.

[17] J.J. Solberg, Stochastic modeling of large scale transportation networks, Rept. No. DOT-ATC-79-2, School of Industrial Engineering, Purdue Univ., West Lafayette, IN, 1979.

[18] K.E. Stecke, Formulation and solution of nonlinear integer production planning problems for flexible manufacturing systems, Management Sci. 29 (3) (1983) 273–288.

[19] K.E. Stecke, Useful models to address FMS operating problems, in: J. Browne and E. Szelke, eds., Proc. Advances in Production Management Systems Conf. Budapest, Hungary (Elsevier Science Publishers B.V.—North-Holland, Amsterdam, 1985) 271–283.

[20] K.E. Stecke, A hierarchical approach to solving machine grouping and loading problems of flexible manufacturing systems, Europ. J. Oper. Res. 24 (3) (1986) 369–378.

[21] K.E. Stecke and T.L. Morin, The optimality of balancing workloads in certain types of flexible manufacturing systems, Europ. J. Oper. Res. 20 (1) (1985) 68–82.

[22] K.E. Stecke and B.W. Schmeiser, Equivalent representa-

tions of system throughput in closed queueing network models of multiserver queues, Working Paper No. 324, Division of Research, Graduate School of Business Administration, The University of Michigan, Ann Arbor, MI, December 1982.

[23] K.E. Stecke and J.J. Solberg, The optimality of unbalancing both workloads and machine group sizes in closed queueing networks of multiserver queues, Oper. Res. 33 (4) (1985) 882–910.

[24] R. Suri, Robustness of queueing network formulas, J. Assoc. Comput. Mach. 30 (3) (1983) 564–594.

[25] K.S. Trivedi and R.E. Kinicki, A mathematical model for computer system configuration planning, in: D. Ferrari, ed., Performance of Computer Installations (North-Holland, Amsterdam, 1978).

[26] K.S. Trivedi and T.M. Sigmon, Optimal design of linear storage hierarchies, J. Assoc. Comput. Mach. 28 (2) (1981) 270–288.

[27] K.S. Trivedi and R.A. Wagner, A decision model for closed queueing networks, IEEE Trans. Software Engrg. 5 (4) (1979) 328–332.

[28] K.S. Trivedi, R.A. Wagner and T.M. Sigmon, Optimal selection of CPU speed, device capabilities and file assignments, J. Assoc. Comput. Mach. 27 (3) (1980) 457–473.

[29] D.D.W. Yao, Some properties of the throughput function of closed networks of queues, Tech. Rept., Dept. of Industrial Engineering, Columbia Univ., New York, 1984.