

Characteristics of optimal workload allocation for closed queueing networks

Heungsoon Felix Lee

Department of Industrial Engineering, Southern Illinois University, Edwardsville, IL 62026-1802, USA

Mandyam M. Srinivasan and Candace A. Yano

Department of Industrial & Operations Engineering, University of Michigan, Ann Arbor, MI 48109-2117, USA

Received 11 December 1989

Revised 25 October 1990

Abstract

Lee, H.F., M.M. Srinivasan and C.A. Yano, Characteristics of optimal workload allocation for closed queueing networks, *Performance Evaluation* 12 (1991) 255–268.

We consider the problem of allocating a given workload among the stations in a multi-server product-form closed queueing network to maximize the throughput. We first investigate properties of the throughput function and prove that it is pseudoconcave for some special cases. Some other characteristics of the optimal workload and its physical interpretation are also provided. We then develop two computational procedures to find the optimum workload allocation under the assumption that the throughput function is pseudoconcave in general. The primary advantage of assuming pseudoconcavity is that, under this assumption, satisfaction of first order necessary conditions is sufficient for optimality. Computational experience with these algorithms provides additional support for the validity of this assumption. Finally, we generalize the solution procedure to accommodate bounds on the workloads at each station.

Keywords: closed queueing networks, optimal allocation, pseudoconcavity, Brouwer's fixed point, multiple servers.

1. Introduction

Closed queueing network (CQN) models are widely used in the modeling and analysis of computer systems and flexible manufacturing systems. A performance measure of interest is the throughput of the system, which is defined as the expected number of job completions by the system per unit time. For analytic tractability, typically the Product-form (PF) assumption [5] is used. Under the PF assumption, the only system parameters required to specify the network with a given number of stations are: (i) the number of customer classes and the population of each, (ii) the mean service time demand (or workload) at a station for each customer class, and (iii) the service rate function at each station [2].

Even under the PF assumption, however, the throughput is a complex, nonlinear function of the

system parameters. The study of the mathematical properties of the throughput function is of interest both in the performance evaluation of a system for system parameters, as well as in the prescription of the system parameters that maximize throughput. We are interested in obtaining some characteristics of this function which enables the search for an optimal allocation of a given workload among the stations in CQNs with multiple servers at each station (the multi-server CQN). We assume the CQN satisfies Product-form and that it has a single class of customers.

The throughput function has been well studied in the case of CQNs with a single server at each station (the single-server CQN). Price [8] shows that the reciprocal of the throughput function is a convex function of the workloads. Shanthikumar and Yao [18], using a sample-path based approach, show that this property holds for a more

general class of non-product-form cyclic queueing networks with single server stations. Under the constraint that a given workload is allocated among the stations of a single-server CQN, Secco-Suardo [11] and Solberg [19] conjecture that the throughput is, in fact, a concave function of the workloads. However, Stecke [21] shows that it is not concave but strictly quasiconcave for a two station CQN, and provides computational evidence that it is strictly quasiconcave for a CQN with more than two stations.

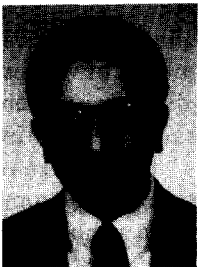
Based on the result of Price, several results have been reported [6,25–28] on the allocation of the workloads that optimize the throughput under various constraints for a central server CQN consisting only of single-server stations. For a given

workload, in the absence of any constants, it has been shown [12,30] that balancing the workload allocated to each station maximizes the throughput in the case of a CQN with only single-server stations.

Yao and Kim [29] prove that balancing the workload maximizes the throughput for a multi-server CQN, when each station has the same number of identical servers. Shanthikumar [13] extends this result to CQNs where the stations have identical, but more general service rate functions. However, Stecke and Solberg [22] report numerical evidence that when the number of servers at each station is not the same, then the throughput is maximized by a unique unbalanced workload allocation to each station. Based on this observa-



Heungsoon Felix Lee is an assistant professor in the Department of Industrial Engineering at Southern Illinois University at Edwardsville. He holds a Ph.D. from the the University of Michigan, an M.S. from Oklahoma State University, and a B.S. from Hanyang University, Korea. His research is mainly concerned with the design and operational control of production facilities manufacturing discrete items, and he has published in *International Journal of Flexible Manufacturing Systems*, and *Journal of Manufacturing Systems*. He is a member of ORSA and IIE.



Mandyam M. Srinivasan is an assistant professor in the Industrial and Operations Engineering Department at the University of Michigan. He received the Master of Technology degree from the Indian Institute of Technology, Madras, India, a Post-Graduate Diploma in Management (M.B.A.) from the Indian Institute of Management, Bangalore, India, and a Ph.D. in industrial engineering and management sciences from Northwestern University. His professional experience includes over five years of full-time employment in leading automobile manufacturing companies in India. His current research interests are in performance evaluation of computer communication networks and material handling systems. Dr. Srinivasan is a member of ACM and TIMS. His most recent papers have appeared in *Management Science*, *IEEE Journal on Selected Areas in Communication*, *Queueing Systems*, *Performance Evaluation*, and *IIE Transactions*. He serves on the editorial boards of *International Journal of Flexible Manufacturing Systems* and *Production and Operations Management*.



Candace Arai Yano is an associate professor in the Department of Industrial and Operations Engineering at the University of Michigan. She holds an A.B. in economics, a M.S. in operations research, and a M.S. and Ph.D. in industrial engineering from Stanford University. Prior to joining the University of Michigan, she was a Member of the Technical Staff at Bell Telephone Laboratories. Her primary research interests are production planning, inventory control, scheduling, the production-quality interface, and logistics. She serves on the editorial boards of *IIE Transactions*, *Interfaces*, *Journal of Manufacturing and Operations Management* and *Naval Research Logistics*, and has authored articles in these journals, *Operations Research*, *Management Science*, and others.

tion, Stecke [20] provides a heuristic algorithm to find an optimal allocation but does not report computational results.

The behavior of the throughput as a function of the number of customers in the CQN has been well studied. It is shown [15,16,24], that the throughput is monotonic increasing with the number of customers in CQNs when every station has a service rate that is concave, non-decreasing with the number of customers present at the station. Shanthikumar and Yao [18] further show that the throughput is a concave function of the number of customers in the CQN.

In this article, we consider the problem of finding the allocation of a given workload among the stations in a multi-server CQN which maximizes the throughput. We refer to this as the *workload allocation problem*. The motivation for this problem is provided in the studies of optimal machine grouping and workload allocation in flexible manufacturing systems [20,22].

The remainder of this article is organized as follows. In Section 2, the multi-server CQN model is defined and the nonlinear programming formulation of the workload allocation problem is stated. In Section 3, we state the Kuhn–Tucker necessary conditions and derive some characteristics that the optimal workload allocation possesses. We then prove pseudoconcavity of the throughput function for some special cases and conjecture that the throughput is pseudoconcave for a general multi-server CQN. Two algorithms to find a workload allocation satisfying the necessary conditions are presented in Section 4 and computational experience with these algorithms is described, for CQNs with numerous combinations of parameter values. Section 5 explains how the solution procedures can be adapted to problems with bounds on the workloads at the various stations. Section 6 concludes with a brief summary.

2. The mathematical formulation

The CQN that we consider consists of M arbitrarily connected multi-server stations, with N customers in the system. The servers at each station are assumed to be identical in terms of their processing capability, and we let S_i , $i = 1, \dots, M$, denote the number of servers at station i . There is a total mean workload, TW , that is to be allocated

among these M stations. Let the workload assignment be denoted by $\bar{W} = (W_1, \dots, W_M)$, where W_i denotes the workload assigned to station i . The workload W_i is the mean service time demanded from station i by a customer in a typical cycle, and is the product of the mean number of visits, v_i that a customer makes to station i in the cycle and the mean service time, τ_i , required per visit, namely, $W_i = v_i \tau_i$. When there are j customers at station i , they are processed at a rate $\mu_i(j)$, where $\mu_i(j) = \min(j, S_i)$. The throughput and the cycle time of the CQN are denoted by $TH(N, \bar{W})$ and $C(N, \bar{W})$ respectively.

Let $G(N, \bar{W})$ denote the normalizing constant for this network. This is defined as

$$G(N, \bar{W}) = \sum_{n_1 + \dots + n_M = N} \prod_{i=1}^M \prod_{j=0}^{n_i} f_i(j), \quad (1)$$

where n_i denotes the number of customers at station i , and $f_i(j)$ is given as:

$$f_i(j) = 1; \quad j = 0, \\ = \frac{W_i}{\mu_i(j)}, \quad j > 0.$$

The throughput of the CQN is given in terms of the normalizing constants as

$$TH(N, \bar{W}) = \frac{G(N-1, \bar{W})}{G(N, \bar{W})}. \quad (2)$$

The performance measures of the CQN, including the throughput, can be obtained for a given set of input parameters using computational algorithms such as the convolution algorithm [4] or the mean value analysis (MVA) algorithm [9], with time complexity $O(MN^2)$.

2.1. Problem formulation

The goal of the workload allocation problem is to allocate the given total mean workload TW among the M stations such that the throughput is maximized. The problem may be mathematically stated as follows:

$$\mathbf{P:} \quad \text{Maximize} \quad TH(N, \bar{W}) \\ \text{subject to:} \quad \sum_{i=1}^M W_i = TW, \quad (3)$$

$$W_i \geq 0, \quad i = 1, \dots, M. \quad (4)$$

3. Characterization of the optimal workloads

The Kuhn–Tucker necessary conditions for Problem P are given by

$$\begin{aligned} \text{KT: } \frac{\partial}{\partial W_i} \text{TH}(N, \bar{W}) + \nu + \pi_i &= 0, \quad i = 1, \dots, M, \\ \sum_{i=1}^M W_i &= \text{TW}, \\ \bar{\pi} &\geq 0, \quad \bar{W} > 0, \quad \bar{\pi} \bar{W} = 0, \end{aligned}$$

where $\partial/\partial W_i$ ($\text{TH}(N, \bar{W})$) is the i th element of the gradient vector for $\text{TH}(N, \bar{W})$ evaluated at \bar{W} , and ν and $\bar{\pi}$ are Lagrange multipliers corresponding to the total workload and workload non-negativity constraints, respectively. The term $\partial/\partial W_i$ ($\text{TH}(N, \bar{W})$) can be expressed as (see, e.g., [6]):

$$\begin{aligned} \frac{\partial}{\partial W_i} \text{TH}(N, \bar{W}) \\ = - \frac{\text{TH}(N, \bar{W})}{W_i} (Q_i(N, \bar{W}) \\ - Q_i(N-1, \bar{W})), \end{aligned} \quad (5)$$

where $Q_i(N, \bar{W})$ is the mean number of customers at station i , with N customers in the CQN.

Consider a CQN where some station, say, k is a delay station. No queueing takes place at station k and hence, if all the workload is assigned to station k , the cycle time is just TW , and the throughput is obviously maximized. The optimum solution \bar{W}^* is thus $W_k^* = \text{TW}$, and $W_j^* = 0$ for $j \neq k$. Lemma 1 shows that this allocation satisfies the necessary conditions given by KT. A proof of Lemma 1 appears in the Appendix.

Lemma 1. *If station k is a delay station, then the optimum solution \bar{W}^* is $W_k^* = \text{TW}$, and $W_j^* = 0$ for $j \neq k$. This solution satisfies the necessary conditions given by KT.*

However, if there is no delay station in the CQN, then any workload allocation which has at least one $W_i = 0$ cannot satisfy KT, as Theorem 1 shows. Theorem 1 requires the following intuitive result, which is stated as Lemma 2. A proof of Lemma 2 appears in the Appendix.

Lemma 2. *If all the stations in the CQN have service rates which are concave and non-decreasing*

with the number of customers present at the station, then the cycle times are non-decreasing in N , i.e., $C(N+1, \bar{W}) \geq C(N, \bar{W})$.

Theorem 1. *If none of the servers are delay stations, then any workload allocation \bar{W} which has at least one $W_k = 0$ cannot satisfy KT; that is, the optimal workload $\bar{W}^* > 0$.*

Proof. Suppose that \bar{W} with $W_k = 0$ for some k satisfies KT. Let $I = \{i \mid W_i > 0\}$ and $\bar{I} = \{k \mid W_k = 0\}$. Clearly, \bar{I} is not empty. From KT, $\pi_i = 0$ and $\text{TH}(N, \bar{W})(Q_i(N, \bar{W}) - Q_i(N-1, \bar{W})) = \nu W_i$, $i \in I$. Summing this equation over all $i \in I$, we have $\nu = \text{TH}(N, \bar{W})/\text{TW}$. It can be easily shown that (also refer to the proof of Lemma 1)

$$\begin{aligned} \lim_{w_k \rightarrow 0} \frac{\partial \text{TH}(N, \bar{W})}{\partial W_k} \\ = - \text{TH}(N, \bar{W})(\text{TH}(N, \bar{W}) \\ - \text{TH}(N-1, \bar{W})); \quad k \in \bar{I}. \end{aligned} \quad (6a)$$

Hence, from KT and Eq. (6a), substituting $\nu = \text{TH}(N, \bar{W})/\text{TW}$, we obtain

$$\begin{aligned} \pi_k = \text{TH}(N, \bar{W}) \{ \text{TH}(N, \bar{W}) \\ - \text{TH}(N-1, \bar{W}) - 1/\text{TW} \}; \\ k \in \bar{I}. \end{aligned} \quad (6b)$$

We now show that $\pi_k < 0$ for each $k \in \bar{I}$, which will imply that \bar{W} does not satisfy KT. When there is no delay server in the network, $C(N, \bar{W}) > \text{TW}$ since there is a positive probability that some customers must wait before being served. Rewriting the term within the braces in Eqn. (6b) using Little's law, we have

$$\begin{aligned} \text{TH}(N, \bar{W}) - \text{TH}(N-1, \bar{W}) - 1/\text{TW} \\ = N/C(N, \bar{W}) - (N-1)/C(N-1, \bar{W}) \\ - 1/\text{TW} \\ = 1/C(N, \bar{W}) - 1/\text{TW} + (N-1) \\ \times (1/C(N, \bar{W}) - 1/C(N-1, \bar{W})). \end{aligned}$$

In the above equation, $1/C(N, \bar{W}) - 1/\text{TW} < 0$. Furthermore, $1/C(N, \bar{W}) - 1/C(N-1, \bar{W}) \leq 0$ from Lemma 2. Thus, $\pi_k < 0$ for each $k \in \bar{I}$ and the result follows. \square

Suppose that $\bar{W} > 0$ satisfies KT. Then $\bar{\pi} = 0$, and (\bar{W}, ν) is the solution to the following system of equations:

$$\frac{\partial}{\partial W_i} \text{TH}(N, \bar{W}) + \nu = 0, \quad i = 1, \dots, M, \quad (7a)$$

$$\sum_{i=1}^M W_i - TW = 0. \tag{7b}$$

Multiplying both eqns. (5) and (7a) by W_i , and summing over all i , we get

$$TH(N, \bar{W}) = \nu TW, \tag{8}$$

and so from eqns. (7a) and (8),

$$\frac{\partial}{\partial W_i} TH(N, \bar{W}) = -TH(N, \bar{W})/TW. \tag{9}$$

From eqns. (5) and (9) we obtain the following expression, which characterizes the number of jobs at each workstation for an allocation which satisfies the necessary conditions for optimality:

$$TW(Q_i(N, \bar{W}) - Q_i(N - 1, \bar{W})) = W_i. \tag{10}$$

To offer some intuition for eqn. (10), we first rewrite this equation as

$$Q_i(N, \bar{W}) - Q_i(N - 1, \bar{W}) = \frac{W_i}{TW},$$

$$i = 1, \dots, M.$$

The above equation states that the optimal workload has the property that when the N th customer is added to the CQN, the mean queue length at each station strictly increases ($\bar{W} > 0$) and the amount of the increase is the same as the ratio of the workload at the station to the total workload. Note that $Q_i(N, \bar{W}) \geq Q_i(N - 1, \bar{W})$ [15,16,23].

Let $g_i(\bar{W}) = TW(Q_i(N, \bar{W}) - Q_i(N - 1, \bar{W}))$. Then, eqn. (10) can be rewritten as

$$g(\bar{W}) = \bar{W} \tag{11}$$

where $g(\bar{W}) = (g_1(\bar{W}), \dots, g_M(\bar{W}))$. In the following, we derive the form of $g(\bar{x})$.

Definition 1. Let g be a continuous function such that $g: \Phi \rightarrow \Phi$, where $\Phi \subset R^n$ is a convex and compact set. Then there exists an $\bar{x} \in \Phi$ such that $g(\bar{x}) = \bar{x}$ by Brouwer's Theorem [24]. This \bar{x} is called a Brouwer's fixed point.

Lemma 3. *The workload \bar{W} satisfying eqn. (11) is a Brouwer's fixed point.*

Proof. Let Γ be the feasible region for problem P. Observe that Γ is an $(M - 1)$ dimensional simplex. Let the function g be defined as the mapping in eqn. (11). Clearly Γ is a convex and compact set, and g is continuous over Γ , since $Q_i(N, \bar{W})$ is continuous over Γ for any nonnega-

tive integer N . Hence, Γ and g satisfy Definition 1. In order to show that $g: \Gamma \rightarrow \Gamma$, we will show that for any $\bar{W} \in \Gamma$, $g(\bar{W}) \in \Gamma$. From results on the monotonicity of queue lengths (with respect to N) for CQNs with concave, non-decreasing service rates at all stations [15,16,23], we have $g_i(\bar{W}) \geq 0$, for all i . Summing g_i over all i , we have

$$\begin{aligned} & \sum_{i=1}^M g_i(\bar{W}) \\ &= \sum_{i=1}^M TW(Q_i(N, \bar{W}) - Q_i(N - 1, \bar{W})) \\ &= TW(N - (N - 1)) = TW. \end{aligned}$$

Thus $g(\bar{W}) \in \Gamma$, and we have shown that \bar{W} satisfying eqn. (11) is a Brouwer's fixed point. \square

It may be observed that any allocation with $W_i = TW$, and $W_j = 0$ for $j \neq i$, is a Brouwer's fixed point satisfying eqn. (11). Lemma 4 formalizes this statement.

Lemma 4. *Every extreme point of Γ is a Brouwer's fixed point satisfying eqn. (11). Hence there are at least M Brouwer's fixed points satisfying eqn. (11) over the feasible region Γ .*

As noted earlier, however, note that an extreme point of Γ may not be a solution for KT.

Remark. The results of Lemmas 1 to 4 apply to a class of product-form CQNs which is more general than the multi-server CQN. Lemma 1 holds for any product-form CQN while Lemmas 2 through 4 hold for CQNs where all stations have service rates which are concave, non-decreasing with the number of customers present at the station. Note that only the queue length monotonicity property, together with eqn. (10), is required to define a Brouwer's fixed point.

3.1. Pseudoconcavity of throughput

Yao and Kim [29] prove that when a multi-server CQN has the same number of servers at each station, the throughput is a Schur-concave function of workload; i.e., it reverses majorization ordering [7]. Using this property, it is shown that balancing workloads among stations maximizes the throughput for these CQNs. Shanthikumar

[13] generalizes this result to networks where the stations have identical concave, non-decreasing service rate functions.

Stecke [21] shows that the throughput is not concave over the feasible region for a single-server CQN and conjectures that it is strictly quasiconcave. The basis for the conjecture is a proof of quasiconcavity for a single-server CQN with two stations and empirical evidence for a single-server CQN with three stations. Stecke also provides some computational evidence that the function is strictly quasiconcave for a multi-server CQN.

We make a stronger conjecture that the function is pseudoconcave over the feasible region for any multi-server CQN. (Pseudoconcavity implies that the function is strictly quasiconcave but the converse is not true.) The following definition for a pseudoconcave function is due to Bazaraa and Shetty [3, see p. 106].

Definition 2. Let $f: S \rightarrow R^1$, where S is a non-empty open set in R^n , and f is differentiable. The function f is said to be *pseudoconvex* if for each $y, z \in S$ with $\nabla f(y)^T(z - y) \geq 0$, we have $f(z) \geq f(y)$, or equivalently, if $f(z) < f(y)$ then $\nabla f(y)^T(z - y) < 0$. The function f is said to be *pseudoconcave* if $-f$ is pseudoconvex.

Property 1. Let $g: \Phi \rightarrow R^1$ and $h: \Phi \rightarrow R^1$, where Φ is a nonempty convex open set in R^n and g is concave, differentiable and nonnegative, and h is convex, differentiable and positive. Then the function f defined by $f(\bar{x}) = g(\bar{x})/h(\bar{x})$ is pseudoconcave. (See 3.41 on page 116 of [3].)

Property 2. If function f is pseudoconcave, then \bar{x} such that $\nabla f(\bar{x}) = 0$ is a global maximum of f . (See page 106 of [3].)

Property 2 is not shared by differentiable strongly or strictly quasiconcave functions. Thus, Stecke's conjecture does not provide a theoretical ground for global optimality of a solution satisfying the necessary conditions. The benefit of assuming pseudoconcavity of the throughput function is that satisfaction of the first order conditions is both necessary and sufficient for optimality in the workload allocation problem.

Note that a CQN with only delay stations, that is, a CQN with $S_i \geq N$ for all i is pseudoconcave in \bar{W} since $TH(N, \bar{W}) = N/TW$ for any \bar{W} in this

CQN. We show that the conjecture is true for two other special cases: the single-server CQN, and the multi-server CQN with $N = 2$. Note that the latter case represents a CQN with only single servers and delay servers.

Let $\Gamma = \{\bar{W}: \sum_i W_i = TW; W_i \geq 0 \text{ for all } i\}$ denote the feasible region for problem P. Since the reciprocal of the throughput function for a CQN with single server stations is convex [8], noting that $TH(N, \bar{W}) = N/C(N, \bar{W})$ and using Property 1, we obtain Lemma 5:

Lemma 5. $TH(N, \bar{W})$ is pseudoconcave over Γ for a single-server CQN. \square

Lemma 6. $TH(N, \bar{W})$ is pseudoconcave over Γ for a multi-server CQN when $N = 2$.

The proof of Lemma 6 is straightforward and is omitted. These lemmas lead to the following conjecture:

Conjecture 1. $TH(N, \bar{W})$ is pseudoconcave over Γ for a multi-server CQN. \square

Note, from Property 1, that a sufficient condition for Conjecture 1 to hold is that $C(N, \bar{W})$ is convex for a multi-server CQN. Empirical support for this conjecture is provided in the following sections. Conjecture 1 and Theorem 1 state that the workload $\bar{W} > \mathbf{0}$ satisfying eqn. (11) is the optimum solution for P. In the following section, we develop two heuristic algorithms to obtain the optimal workload allocation under the pseudoconcavity assumption: the Eaves-Saigal fixed point algorithm and the reduced gradient algorithm. We also report on the performance of these algorithms.

4. The solution procedure

We coded two algorithms, the reduced gradient algorithm and the Eaves-Saigal fixed point algorithm, to solve Problem P. Both algorithms use as an initial feasible point a balanced allocation (i.e., the total mean workload is allocated such that the W_i/S_i ratios are equal). Both procedures search the feasible region systematically in order to improve the throughput while maintaining feasibility. Both terminate at a point which satisfies the necessary conditions.

Stecke [20] gives a sketch of an algorithm for this workload allocation problem but does not report computational results. We expect that our algorithms are more efficient for two reasons. First, the algorithm proposed by Stecke requires a line search to be performed at each iteration, varying only two W_i s with the remaining W_j s fixed. Second, her algorithm requires the computation of M throughputs (to provide approximate partial derivative information) to determine which two W_i s are to be varied. Her algorithm terminates when the sensitivity information indicates that further workload changes cannot increase the throughput. We explain below how our procedures differ.

The reduced gradient algorithm [1] uses reduced gradient vectors by eliminating the dependent variables from the equality constraint, that is, eqn. (3). At each iteration, a steepest ascent direction is derived in the space of independent variables and a line search is performed along the direction. Thus all W_i s can change at each iteration. Calculating a reduced gradient vector does not require any extra computation since an entire gradient vector is obtained with only one throughput calculation using the MVA algorithm.

In order to apply the Eaves–Saigal fixed point algorithm, Problem P is equivalently rewritten as

$$\begin{aligned} \text{P}' : \quad & \text{Minimize} \quad \theta(\bar{W}) = -\text{TH}(N, \bar{W}), \\ & \text{subject to} \quad s(\bar{W}) \leq 0, \end{aligned}$$

where $\bar{W} = (W_1, \dots, W_{M-1}, \text{TW} - \sum_{i=1}^{M-1} W_i)$, and $s(\bar{W}) = \max\{\sum_{i=1}^{M-1} W_i - \text{TW}, [\max_i(-W_i), i = 1, \dots, M-1]\}$.

Now, define the following point-to-set mapping $p(\bar{W})$ as

$$p(\bar{W}) = \begin{cases} \nabla\theta(\bar{W}); & \text{if } s(\bar{W}) < 0 \\ \text{the convex hull of} \\ \quad \nabla\theta(\bar{W}) \text{ and } \nabla s(\bar{W}); & \text{if } s(\bar{W}) = 0 \\ \nabla s(\bar{W}); & \text{if } s(\bar{W}) > 0, \end{cases} \quad (12)$$

where $\nabla\theta(\bar{W})$ and $\nabla s(\bar{W})$ are the gradient vectors of $\theta(\bar{W})$ and $s(\bar{W})$, respectively. It can be shown that the point \bar{W} satisfying the conditions $\mathbf{0} \in p(\bar{W})$, and $s(\bar{W}) \leq 0$, satisfies the necessary conditions KT.

Theorem 2. If $N \geq \max_i(S_i)$ and there exists \bar{W} such that $\mathbf{0} \in p(\bar{W})$, then the Eaves–Saigal algorithm converges to it quadratically.

Proof. If $N > \max_i(S_i)$, then $\bar{W} > \mathbf{0}$ from Theorem 1. This implies $s(\bar{W}) < 0$; that is, it is not a point on the boundary. Thus, $\mathbf{0} \in p(\bar{W})$ can be stated equivalently as $\mathbf{0} = \nabla\theta(\bar{W})$. This workload \bar{W} satisfies eqn. (11), and is a Brouwer’s fixed point from Lemma 3. Therefore, the algorithm converges to it quadratically [20]. \square

The Eaves–Saigal algorithm appears to be the only algorithm with the property of quadratic converge for this problem. Since the (reduced) Hessian matrix will not be negative definite due to the nonconcavity of the function, any Newton-type method requires a line search to be performed, which only allows linear convergence. Also, each calculation of the Hessian requires the computation of $O(M)$ throughputs. This is because the mean queue lengths must be evaluated for the CQN, Ψ , with M stations, and mean queue lengths must also be evaluated for M other CQNs, $\Psi^{(i)}$, $i = 1, \dots, M$, each one of which is identical to CQN Ψ , but with station i removed.

4.1. Experimental results

We conducted a number of experiments using the two algorithms described above, for CQNs with a range of parameter values. In these experiments, the number of stations, M , ranged from 2 to 10. Arbitrary unbalanced configurations for the server vector \bar{S} were chosen, and the number of customers, N ranged from $(\max_i(S_i) + 1)$ to 50. We chose N to be greater than $\max_i(S_i)$ since otherwise a trivial optimal solution is available from Lemma 1. The workloads were scaled such that $\text{TW} = \sum_{i=1}^M S_i$ without loss of generality [22]. We used the following as a termination condition:

$$D(\bar{W}) = \max_i |W_i - \text{TW}(Q_i(N, \bar{W}) - Q_i(N-1, \bar{W}))| \leq \epsilon, \quad (13)$$

for some specified tolerance ϵ . The following statistics were collected at termination: throughput, the number of throughput computations, and the two-norm of the steepest ascent direction. The last statistic was collected in order to indicate the slope of the throughput function at the point of

Table 1
Reduced gradient algorithm vs Eaves–Saigal algorithm at $N = 5$

System configuration	Algorithm	Throughput	# Throughput computations	Two-norm of steepest ascent feasible direction
$M = 2$ $\bar{S} = (1, 3)$	reduced gradient	0.84218	9	1×10^{-4}
	Eaves–Saigal	0.84219	8	2×10^{-12}
$M = 3$ $\bar{S} = (1, 2, 4)$	reduced gradient	0.65392	58	4×10^{-7}
	Eaves–Saigal	0.65392	10	1×10^{-6}
$M = 4$ $\bar{S} = (2, 2, 2, 4)$	reduced gradient	0.48059	21	3×10^{-7}
	Eaves–Saigal	0.48059	14	8×10^{-10}
$M = 5$ $\bar{S} = (1, 3, 3, 3, 4)$	reduced gradient	0.35417	233	2×10^{-7}
	Eaves–Saigal	0.35416	16	1×10^{-6}
$M = 6$ $\bar{S} = (1, 2, 2, 3, 4, 4)$	reduced gradient	0.31093	243	4×10^{-7}
	Eaves–Saigal	0.31094	60	8×10^{-7}
$M = 7$ $\bar{S} = (1, 2, 2, 3, 4, 4)$	reduced gradient	0.27610	352	3×10^{-7}
	Eaves–Saigal	0.27610	41	7×10^{-8}
$M = 8$ $\bar{S} = (1, 1, 1, 1, 1, 1, 4)$	reduced gradient	0.41555	226	2×10^{-7}
	Eaves–Saigal	0.41555	42	9×10^{-8}

termination. These statistics are summarized in Tables 1 and 2 below, for $\varepsilon = 0.01$.

Over a 100 configurations were tested. For every problem, both the algorithms always converged to the same interior point. We also tried

four different initial points which were randomly generated from the feasible region. For every initial point, both algorithms still converged to the same interior point. This observation leads to the following conjecture.

Table 2
Reduced gradient algorithm vs Eaves–Saigal algorithm at $N = 20$

System configuration	Algorithm	Throughput	# Throughput computations	Two-norm of steepest ascent feasible direction
$M = 2$ $\bar{S} = (1, 3)$	reduced gradient	0.95997	9	5×10^{-6}
	Eaves–Saigal	0.95997	10	3×10^{-10}
$M = 3$ $\bar{S} = (1, 2, 4)$	reduced gradient	0.91374	30	3×10^{-6}
	Eaves–Saigal	0.91374	19	8×10^{-8}
$M = 4$ $\bar{S} = (2, 2, 2, 4)$	reduced gradient	0.85599	56	9×10^{-7}
	Eaves–Saigal	0.85599	13	3×10^{-9}
$M = 5$ $\bar{S} = (1, 2, 3, 4, 7)$	reduced gradient	0.79851	151	9×10^{-8}
	Eaves–Signal	0.79851	25	3×10^{-8}
$M = 6$ $\bar{S}(1, 2, 2, 3, 6, 8)$	reduced gradient	0.73428	232	1×10^{-8}
	Eaves–Saigal	0.73428	32	9×10^{-7}
$M = 7$ $\bar{S} = (1, 2, 2, 3, 3, 4)$	reduced gradient	0.72300	179	8×10^{-8}
	Eaves–Saigal	0.72300	25	5×10^{-8}
$M = 8$ $\bar{S} = (1, 1, 2, 2, 3, 3, 5, 9)$	reduced gradient	0.65927	274	5×10^{-9}
	Eaves–Saigal	0.65927	60	6×10^{-8}

Conjecture 2. If $N > \max_i(S_i)$, then there is a unique solution, \bar{W} , for KT.

The solution \bar{W} is globally optimal under *either* Conjecture 1 *or* Conjecture 2, and from Theorem 2 it can be found by the Eaves–Saigal algorithm with quadratic convergence. For small values of ϵ (i.e., $\epsilon < 1 \times 10^{-2}$), the Eaves–Saigal algorithm requires a far smaller number of throughput computations than the reduced gradient algorithm. The reduced gradient algorithm was not executed for $\epsilon < 1 \times 10^{-2}$ due to its slow convergence. Note that our conjecture that the solution is unique is consistent with the observation made by Stecke and Solberg [22] and Stecke [20] that there is a unique unbalanced optimal allocation for Problem P.

Conjecture 2 and Lemma 4 lead to Proposition 1, a proof of which appears in the Appendix.

Proposition 1. If $N > \max_i(S_i)$, then there are exactly $2^M - 1$ Brouwer’s fixed points satisfying eqn. (11) over the feasible region, Γ , when Conjecture 2 holds.

5. Generalization to include workload bounds

In this section, we generalize the solution procedures to accommodate lower and upper bounds on the workload at each station. These bounds might arise for a variety of different reasons. For example, upper bounds might be specified to allow adequate time for planned maintenance and lower bounds can ensure a minimum level of machine utilization. The problem is the same as P, but the non-negativity constraints for W_i are replaced by constraints of the form $L_i \leq W_i \leq U_i$ where L_i and U_i are the lower and upper bounds, respectively, on W_i . Note that when there are bounds on the workload allocations, even the case where there are delay servers present in the network need not be trivial. However, for this case, we can allocate the maximum possible load among the delay servers, and solve the workload allocation problem to allocate the remaining workload (if any) among the remaining stations.

One important difference between the constrained problem and the more general one is that a balanced workload allocation may not be feasible for the constrained problem. We assume that

there is a feasible workload allocation (i.e., $\sum_i L_i \leq \text{TW} \leq \sum_i U_i$).

The following algorithm is used to find a good initial feasible solution. Let Ω denote the set of stations; initially, $\Omega = \{1, \dots, M\}$. Throughout the algorithm, A denotes the set of stations for which the workload exceeds the upper bound, B denotes the set of stations for which the workload is less than the lower bound and C denotes the set of stations whose workloads are within bounds. In other words, $A = \{i \mid W_i > U_i, i \in \Omega\}$, $B = \{i \mid W_i < L_i, i \in \Omega\}$, and $C = \Omega - A - B$, where W_i is the workload currently allocated to station i . Let $S_A = \sum_{i \in A} (W_i - U_i)$ and $S_B = \sum_{i \in B} (L_i - W_i)$.

Algorithm WB

1. Find a balanced workload allocation $\bar{W} = \{W_i\}$. If it is feasible then **stop**. Otherwise, construct the sets A , B , and C . Compute S_A , S_B , and let $\Delta = S_A - S_B$. If $\Delta > 0$, go to step 2; otherwise if $\Delta < 0$, go to step 3. If $\Delta = 0$, reduce the workload allocated to the stations in A by an amount S_A , allocate S_B among the stations in B , and **stop**.
2. a) Reset the workloads for all stations in the set A at their upper bounds, and update $\Omega = \Omega - A$. Allocate Δ equally among all the stations in the set $B \cup C$, and let $S_B^{\text{old}} = S_B$.
 b) Update A , B , C , and reevaluate S_A and S_B . Compute $\Delta = (S_A + S_B^{\text{old}}) - S_B$.
 c) If $S_B^{\text{old}} = S_B$, allocate S_B among the stations in B and **stop**; otherwise go to step 2a).
3. a) Reset the workloads for all stations in the set B at their lower bounds, and update $\Omega = \Omega - B$. Allocate Δ equally among all the stations in the set $A \cup C$, and let $S_A^{\text{old}} = S_A$.
 b) Update A , B , C , and reevaluate S_A and S_B . Compute $\Delta = S_A - (S_A^{\text{old}} + S_B)$.
 c) If $S_A^{\text{old}} = S_A$, reduce S_A from the stations in A and **stop**; otherwise go to step 3a).

At the end of Algorithm WB we can identify three sets of stations: a) the set X in which each station has its workload at the upper bound, b) the set Y in which each station has its workload at the lower bound, and c) the set Z , which consists of the rest of the stations each of which has a workload strictly within its bounds. Let \bar{W} denote the allocation obtained from the algorithm. When every station has the same number of servers, these workloads have the following property:

- a) $W_i > W_j, i \in Z, j \in X$

- b) $W_i = W_j, i, j \in Z;$
- c) $W_i < W_j, i \in Z, j \in Y.$

It can easily be shown that any reallocation of these workloads between stations will either be infeasible, or will result in a workload vector that “majorizes” \bar{W} . (For a definition of majorization ordering, see [7]). Thus, the algorithm returns the optimal allocation when every station has the same number of servers [13,29]. Lemma 7 formalizes this statement.

Lemma 7. *If $S_1 = \dots = S_M$, then Algorithm WB gives the optimal workload allocation. □*

To understand the algorithm intuitively, suppose, for example, that $S_A - S_B > 0$ in step 1. For each $i \in A$, the algorithm sets $W_i = U_i$, and removes station i from the list of stations being considered. Following this, the algorithm redistributes the workload $S_A - S_B$ equally among the stations in the sets B and C , and updates the sets A, B , and C , as well as the variables S_A and S_B . The new workloads will result in a changed (reduced) value for S_B which implies that the

workload $(S_B^{\text{old}} - S_B)$ will have to be reallocated, where S_B^{old} denotes the previous value of S_B . The workloads for all stations in the set A need to be readjusted so that they are at their upper bounds, and this “frees up” an additional workload S_A which also needs to be reallocated. The “surplus” workload $(S_A + S_B^{\text{old}} - S_B)$ is reallocated among the stations in sets B and C . The sets A, B , and C , and the variables S_A and S_B are updated once more, and the iteration continues.

If at some point in the iteration, the old and new values of S_B coincide, it implies that no workload was reallocated. At this point, we distribute S_B among the stations in set B and stop.

5.1. Experimental results

We conducted a number of experiments using the two nonlinear programming algorithms and Algorithm WB, for CQNs with a wide range of parameter values. The value of M ranged from 2 to 8 and N was fixed at either 5 or 20. We chose arbitrary unbalanced configurations for \bar{S} . Workloads were scaled such that the total

Table 3
Seven sets of Problem P with loose (a) and tight (b) bound constraints

(1) $M = 2, \bar{S} = (3, 1)$	
a) $\bar{L} = (0.1, 0.1)$	$\bar{U} = (4, 3)$
b) $\bar{L} = (2, 1)$	$\bar{U} = (4, 3)$
(2) $M = 3, \bar{S} = (4, 2, 1)$	
a) $\bar{L} = (0.5, 0.5, 0.5)$	$\bar{U} = (6, 6, 6)$
b) $\bar{L} = (1, 3, 1)$	$\bar{U} = (5, 5, 5)$
(3) $M = 4, \bar{S} = (4, 2, 2, 2)$	
a) $\bar{L} = (1, 1, 0.1, 0.1)$	$\bar{U} = (6, 5, 5, 5)$
b) $\bar{L} = (2, 2, 1, 1)$	$\bar{U} = (4, 4, 4, 4)$
(4) $M = 5, \bar{S} = (7, 4, 3, 2, 1)$	
a) $\bar{L} = (0.9, 0.8, 0.7, 0.6, 0.5)$	$\bar{U} = (14, 13, 13, 12, 12)$
b) $\bar{L} = (2, 2, 2, 2, 2)$	$\bar{U} = (10, 10, 10, 10, 10)$
(5) $M = 6, \bar{S} = (8, 6, 3, 2, 2, 1)$	
a) $\bar{L} = (1, 0.9, 0.8, 0.7, 0.6, 0.5)$	$\bar{U} = (18, 18, 18, 15, 15, 15)$
b) $\bar{L} = (3, 2, 3, 3, 2, 1)$	$\bar{U} = (15, 15, 15, 10, 10, 10)$
(6) $M = 7, \bar{S} = (4, 3, 3, 3, 2, 2, 1)$	
a) $\bar{L} = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.3)$	$\bar{U} = (6, 6, 6, 6, 6, 6, 6)$
b) $\bar{L} = (3, 3, 3, 1, 1, 1, 2)$	$\bar{U} = (4, 4, 4, 4, 4, 4, 4)$
(7) $M = 8, \bar{S} = (9, 5, 3, 3, 2, 2, 1, 1)$	
a) $\bar{L} = (0.5, 0.5, 0.5, 0.5, 0.5, 0.1, 0.1, 0.1)$	$\bar{U} = (21, 21, 21, 21, 21, 21, 21, 21)$
b) $\bar{L} = (2, 2, 2, 2, 3, 1, 0.5, 0.5)$	$\bar{U} = (20, 20, 20, 20, 20, 20, 20, 20)$

workload, $TW = \sum_{i=1}^M S_i$, without loss of generality. We used two sets of the bounds (loose and tight) for each problem. The bounds were specified so the feasible region for the loose bounds contains the feasible region for the tight ones. Problem data used in the experiment are presented in Table 3.

Since Condition (13) is no longer a valid termination condition, we used the two-norm of the steepest ascent direction, denoted as ∇^2 as the criterion for termination. Clearly, $\bar{W} > 0$ satisfies the Kuhn–Tucker necessary conditions when $\nabla^2 = 0$. The algorithms terminate when $\nabla^2 \leq \epsilon$ for some specified tolerance ϵ . We set $\epsilon = 1 \times 10^{-5}$. Both algorithms were initialized with the solution from Algorithm WB. The following statistics were collected at each termination: throughput, the

number of throughput computations (since throughput calculations take most of the computation time), and the number of active bound constraints. These statistics are summarized in Table 4.

Algorithm WB provides a good initial feasible solution when the bounds are tight. In fact, the initial solution is optimal for all the problems with tight bounds except for problem (7b) with $N = 20$. For problems with no active bound constraint at the optimum, for example, problems with the loose bounds and $N = 20$, the Eaves–Saigal algorithm allows quadratic convergence and requires a smaller number of throughput computations than the reduced gradient algorithm. However, the Eaves–Saigal algorithm converges only linearly for

Table 4
Results for workload allocation problems with workload bounds, with

Problem	Terms	Reduced gradient algorithm		Eaves–Saigal algorithm	
		$N = 5$	$N = 20$	$N = 5$	$N = 20$
(1a)	TH§	0.8422	0.9600	0.8422	0.9600
	no*, act#	31, 0	6, 0	6, 0	8, 0
(1b)	TH	0.7955	0.9497	0.7955	0.9497
	no, act	1, 1	1, 1	1, 1	1, 1
(2a)	TH	0.6509	0.9137	0.6509	0.9137
	no, act	28, 0	44, 0	10, 0	10, 0
(2b)	TH	0.5457	0.6664	0.5457	0.6664
	no, act	1, 2	1, 2	1, 2	1, 2
(3a)	TH	0.4800	0.8560	0.4803	0.8560
	no, act	61, 0	18, 0	11, 0	6, 0
(3b)	TH	0.4662	0.8492	0.4662	0.8492
	no, act	1, 3	1, 3	1, 3	1, 3
(4a)	TH	0.2913	0.7985	0.2925	0.7985
	no, act	8, 1	64, 0	36, 0	19, 0
(4b)	TH	0.2722	0.4994	0.2723	0.4994
	no, act	1, 3	1, 3	1, 3	1, 3
(5a)	TH	0.2255	0.7341	0.2266	0.7342
	no, act	2, 1	67, 0	53, 0	33, 0
(5b)	TH	0.2229	0.6202	0.2229	0.6202
	no, act	1, 4	1, 4	1, 4	1, 4
(6a)	TH	0.2739	0.7229	0.2758	0.7230
	no, act	2, 1	48, 0	57, 0	16, 0
(6b)	TH	0.2588	0.4982	0.2588	0.4982
	no, act	1, 5	1, 5	1, 5	1, 5
(7a)	TH	0.1914	0.6588	0.1909	0.6593
	no, act	3, 1	53, 0	4, 0	51, 0
(7b)	TH	0.1905	0.5985	0.1905	0.5993
	no, act	1, 3	42, 1	1, 3	80, 1

Note: § TH = throughput; no* = the number of throughput computations; and act# = the number of active bound constraints at the termination point

problems with one or more bound constraints active at the optimum solution because of the manner in which it handles constraints.

7. Conclusion

In this article, we considered the problem of allocating a given workload among the stations in a multi-server product-form CQN to maximize the throughput. This problem has been addressed by several researchers. However, there has been relatively little work done in characterizing the optimal workload allocation for a general multi-server CQN and developing an efficient algorithm which exploits these characteristics.

We first stated the nonlinear programming formulation of the workload allocation problem for the multi-server CQN and derived some characteristics that the optimal workload allocation possesses. The optimal workload has the property that when a customer is added to the CQN, the mean queue length at each station strictly increases and the amount of the increase is the same as the ratio of the workload at the corresponding station to the total workload. We showed that if the number of customers in the CQN is greater than the maximum number of servers at any station, the optimal workload is an interior Brouwer’s fixed point when there are no bounds on the workloads.

We then investigated the behavior of the throughput function and proved that it is pseudo-concave for some special cases of the multi-server CQN. The advantage of having a pseudoconcave function is that the Kuhn–Tucker necessary conditions are sufficient for global optimality. Under the pseudoconcavity assumption, we showed that the optimal workload allocation can be found by the Eaves–Saigal fixed point algorithm which has quadratic convergence. We observed numerically that the optimal workload is always unique. Computational experience with algorithms developed to find the optimal workload allocation supports the pseudoconcavity conjecture for the general multi-server CQN. Lastly, we generalized the solution procedure to accommodate bounds on the workloads at each station.

Acknowledgement

We would like to thank Professor Romesh Saigal for permission to use the MP5 software for the Eaves–Saigal fixed point algorithm and for helpful discussions on this topic.

Appendix

Lemma 1. *If station k is a delay station, then the optimum solution \bar{W}^* is $W_k^* = TW$, and $W_j^* = 0$ for $j \neq k$. This solution satisfies the necessary conditions given by KT.*

Proof. When all the workload is assigned to station k , then $TH(N, \bar{W}) = N/TW$. Hence, from eqn. (5) and the fact that $\lim_{W_k \rightarrow TW} Q_k(N, \bar{W}) = N$, we have

$$\begin{aligned} \lim_{W_k \rightarrow TW} \frac{\partial TH(\bar{W})}{\partial W_k} &= -\frac{TH(\bar{W})}{TW} (N - (N - 1)) \\ &= -\frac{TH(\bar{W})}{TW} = -\frac{N}{TW^2}. \end{aligned}$$

As $W_j \rightarrow 0$ for $j \neq k$, the mean sojourn time at station j , $R_j(N, \bar{W}) \rightarrow \tau_j$, where τ_j is the mean service time at station j . This follows, since a customer is very likely to find a server available, and therefore stays at station j only for the duration of a service time. From Little’s law, $Q_j(N, \bar{W}) = v_j TH(N, \bar{W}) R_j(N, \bar{W}) = v_j TH(N, \bar{W}) \tau_j$. Noting that $W_j = v_j \tau_j$, we have

$$\begin{aligned} \lim_{W_j \rightarrow 0} \frac{\partial TH(\bar{W})}{\partial W_j} &= \lim_{W_j \rightarrow 0} -\frac{TH(N, \bar{W})}{W_j} \\ &\quad \times (v_j TH(N, \bar{W}) \tau_j - v_j TH(N - 1, \bar{W}) \tau_j) \\ &= -\frac{N}{TW^2}. \end{aligned}$$

Given the derivative values at \bar{W} , we now solve the necessary conditions. Note that $W_k > 0$ implies $\pi_k = 0$. Thus $\nu = -\partial TH(N, \bar{W})/\partial W_k = N/TW^2$, and $\pi_j = -\partial TH(N, \bar{W})/\partial W_j - \nu = 0$ for each $j \neq k$. Therefore, the values for \bar{W} , $\bar{\pi}$, and ν obtained above provide the solution to KT. □

Lemma 2. *If all the stations in the CQN have service rates which are concave and non-decreasing with the number of customers present at the station, then the cycle times are non-decreasing in N , i.e., $C(N + 1, \bar{W}) \geq C(N, \bar{W})$.*

Proof. Since $C(N, \bar{W}) = \sum_{i=1}^M R_i(N, \bar{W})$, where $R_i(N, \bar{W})$ is the mean sojourn time at station i , we only need to show that $R_i(N + 1, \bar{W}) \geq R_i(N, \bar{W})$ for all i . From eqn. (2.19) of [9], letting $p_i(n|N)$ denote the probability that n customers are present at station i , we have

$$R_i(N, \bar{W}) = \tau_i \left[1 + \frac{1}{S_i} \sum_{n=S_i}^{N-1} p_i(n|N-1) + \frac{1}{S_i} \sum_{n=S_i}^{N-1} (n - S_i) p_i(n|N-1) \right],$$

Since $p_i(N|N-1) = 0$, $\Delta R_i(N, \bar{W}) = R_i(N + 1, \bar{W}) - R_i(N, \bar{W})$ is written as

$$\begin{aligned} \Delta R_i(N, \bar{W}) &= \frac{\tau_i}{S_i} \left[\sum_{n=S_i}^N (p_i(n|N) - p_i(n|N-1)) + \sum_{n=S_i}^N (n - S_i)(p_i(n|N) - p_i(n|N-1)) \right] \\ &= \frac{\tau_i}{S_i} \left[(p_i(n \geq S_i|N) - p_i(n \geq S_i|N-1)) + \sum_{k=S_i+1}^N (p_i(n \geq k|N) - p_i(n \geq k|N-1)) \right] \\ &= \frac{\tau_i}{S_i} \sum_{k=S_i}^N (p_i(n \geq k|N) - p_i(n \geq k|N-1)). \end{aligned}$$

Since $p_i(n \geq k|N) - p_i(n \geq k|N-1) \geq 0$ [15,16,23], each term in the last equation is ≥ 0 . Therefore, $\Delta R_i(N, \bar{W}) \geq 0$ for all i and the result follows. \square

Proposition 1. *If $N > \max_i(S_i)$, then there are exactly $2^M - 1$ Brouwer's fixed points satisfying eqn. (11) over the feasible region of Problem P when Conjecture 2 holds.*

Proof. Consider loading only the first $p \geq 1$ stations in the CQN, fixing $W_i = 0$ for $i = p + 1$ to

M , and solve Problem P for the p -station CQN. Since $N > \max(S_1, \dots, S_p)$, by Conjecture 2 there is only one solution for the necessary conditions of the reduced problem, and this solution (W_1, \dots, W_p) is an interior point. Clearly, $\bar{W} = (W_1, \dots, W_p, 0, \dots, 0)$ is a Brouwer's fixed point satisfying eqn. (11) for the original problem, and there are $\binom{M}{p}$ such Brouwer's fixed points. Therefore, the total number of Brouwer's fixed points satisfying eqn. (11) is given as $\sum_{p=1}^M \binom{M}{p} = 2^M - \binom{M}{0} = 2^M - 1$.

References

- [1] M. Avriel, *Nonlinear Programming Analysis and Methods* (Prentice-Hall, 1976).
- [2] 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** (1975) 248-260.
- [3] M. Bazaraa and C.M. Shetty, *Nonlinear Programming Theory and Algorithms*, (Wiley and Sons, 1979).
- [4] J.P. Buzen, Computational algorithms for closed queueing networks with exponential servers, *Comm. ACM*, **16** (9) (1973) 527-531.
- [5] W.J. Gordon and G.F. Newell, Closed queueing networks with exponential servers, *Oper. Res.* **15** (1967) 252-267.
- [6] H. Kobayashi and M. Gerla, Optimal routing in closed queueing networks, *ACM Trans. Comput. Systems* **1** (4) (1983) 294-310.
- [7] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications* (Academic, 1979).
- [8] T.G. Price, Probability models of multiprogrammed computer systems. Ph.D. dissertation, Department of Electrical Engineering, Stanford University, Stanford CA, 1974.
- [9] M. Reiser and S. Lavenberg, Mean-value analysis of closed multichain queueing networks, *J. Assoc. Comput. Mach.* **27** (2) (1980) 313-322.
- [10] R. Saigal, On the convergence rate of algorithms for solving equations that are based on methods of complementary pivoting, *Math. Oper. Res.* **2** (1977) 108-24.
- [11] G. Secco-Suardo, Optimization of a closed network of queues, Report No. ESL-FR-834-3, Electronic Systems Laboratory, M.I.T., Cambridge, 1978.
- [12] J.G. Shanthikumar, On the superiority of balanced load in a flexible manufacturing system, Technical Report, Dept of IE&OR, Syracuse University, 1982.
- [13] J.G. Shanthikumar, Stochastic majorization of random variables with proportional equilibrium, rates, *Adv. in Appl. Probab.* **19** (1987) 854-872.
- [14] J.G. Shanthikumar and K. Steckel, Reducing work-in-process inventory in certain classes of flexible manufacturing systems, *European J. Oper. Res.*, **26** (1986) 266-271.
- [15] J.G. Shanthikumar and D.D. Yao, The effect of increasing service rates in a closed queueing network, *J. Appl. Probab.*, **23** (1986) 474-483.
- [16] J.G. Shanthikumar and D.D. Yao, Stochastic monotonic-

- ity of the queue lengths in closed queueing networks, *Oper. Res.*, **35** (1987) 583–588.
- [17] J.G. Shanthikumar and D.D. Yao, On server allocation in multiple center manufacturing systems, *Oper. Res.*, **36** (1988) 333–342.
- [18] J.G. Shanthikumar and D.D. Yao, Second-order stochastic properties in queueing systems, *Proc. IEEE*, **77** (1) (1989) 162–170.
- [19] J.J. Solberg, Stochastic modeling of large scale transportation networks, Report No. DOT-ATC-79-2, School of Industrial Engineering, Purdue University, West Lafayette, IN, 1979.
- [20] K.E. Stecke, A hierarchical approach to solving machine grouping and loading problems of flexible manufacturing systems, *European J. Oper. Res.*, **24** (1986) 369–378.
- [21] K.E. Stecke, On the nonconcavity of throughput in certain closed queueing networks, *Performance Eval.*, **6** (4) (1986) 293–305.
- [22] K.E. Stecke and J.J. Solberg, The optimality of unbalancing both workloads and machine group sizes in closed queueing networks of multi-server queues, *Oper. Res.*, **33** (4) (1985) 882–910.
- [23] R. Suri, A concept of monotonicity and its characterization for closed queueing networks, *Oper. Res.* (1984) 606–624.
- [24] M.J. Todd, *The Computation of Fixed Points and Applications* (Springer-Verlag, Berlin-Heidelberg, 1976).
- [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 Eng.* **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. Yao and S.C. Kim, Some order relations in closed networks of queues with multiserver stations, *Naval Res. Logist.* **34** (1987) 53–66.
- [30] J. Zahorjan, K.C. Sevcik, D.L. Eager and B. Galler, Balanced job bound analysis of queueing networks, *Comm. ACM* **25** (1982) 134–141.