Lee, Heungsoon F. (Heungsoon Felix)

SOME CHARACTERISTICS OF OPTIMAL WORKLOAD ALLOCATION FOR CLOSED QUEUEING NETWORKS

Heungsoon F. Lee

M.M. Srinivasan

Candace A. Yano

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

Technical Report 88-9

July 1988

Abstract

In this paper, we consider the problem of allocating a given workload among the stations in a multi-server product form CQN to maximize the throughput. We first investigate properties of the throughput function and prove that it is pseudo-concave in several special cases. We then develop two computational procedures to find the optimum workload allocation under the assumption that the throughput function is pseudo-concave in general. The primary advantage of assuming pseudo-concavity 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. We also show that the optimal workload is an interior Brouwer's fixed point and can be found by the Eaves-Saigal fixed point algorithm which allows quadratic convergence. Some other characteristics of the optimal workload and its physical interpretation are also provided. We also generalize the solution procedure to accommodate bounds on the workloads at each station.

1. INTRODUCTION

Closed queueing network (CQN) models are widely used in the modeling and analysis of computer systems and flexible manufacturing systems. One of the performance measures of interest is the throughput of the system, which is defined as the number of job completions by the system per unit time. For analytic tractability, typically the Product Form (PF) assumption (Gordon and Newell [1967]) is used. Under the PF assumption, the only system parameters required to specify the network with a given number of stations are (Baskett et.al. [1976]) (i) the number of customer classes and their population (ii) the mean service time demand at a station (or workload) for each customer class, and (iii) the service rate function at each station, which is determined by whether the station is a single-server station, a multi-server station, a delay (infinite-server) station, or a station with an arbitrary load dependent service rate function.

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 given the system parameters, as well as in the prescription of the optimal system parameters that maximize the throughput. We are interested, here, in obtaining some characteristics of the throughput 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 make the PF assumption and assume that the CQN 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 [1974] shows that the reciprocal of the throughput function is a convex function of the workloads. This result has also been obtained by Kenevan and Mayrhauser [1984], who show, in addition, that the throughput is a log-convex function of the number of customers in a CQN with an arbitrary number of single-servers and delay servers. Under the constraint that a given workload is allocated among the stations of a single-server CQN, Secco-Suardo [1978] and Solberg [1979] conjectured that the throughput is, in fact, a concave function of the workloads. However, Stecke [1986b] proves that it is not concave but strictly quasiconcave for a 2 station CQN, and provides some computational evidence that it is strictly quasiconcave for a CQN with more than 2 stations.

Based on the result of Price, several results have been reported (Trivedi and Kinicki [1978], Trivedi and Wagner [1979], Trivedi, Wagner and Sigmon [1980], Trivedi and Sigmon [1981], Kobayashi and Gerla [1983]) which enable an allocation of the workloads that optimize the throughput under various constraints for the special case of a central server CQN consisting only of single-server stations. For a given workload, in the absence of any constraints, it can be shown (Yao [1985], Morin and Stecke [1985]) that balancing the workloads allocated to each station maximizes the throughput in the case of a CQN with only single-server stations.

For the multi-server CQN, Yao [1985], and Stecke and Solberg [1985] prove that balancing the workloads maximizes the throughput when each station has the same number of servers. However, Stecke and Solberg graphically observe 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 observation, Stecke [1986a] provides an algorithm which intends to find an optimal allocation but its computational results have not been reported. Yao and Shanthikumar [1986] study the server allocation problem which is to allocate a given number of identical servers among the stations in the CQN to maximize the throughput. They find that the optimal allocation satisfies the 'decreasing property' (more servers are allocated to a station with larger workload) and use this to reduce the number of possible allocations searched.

In this paper, we consider the workload allocation problem which is to find the optimal allocation of a given workload among the stations in a multi-server CQN which maximizes the throughput. The motivation for this problem is provided in the studies of optimal machine grouping and workload allocation in flexible manufacturing systems (Stecke [1983], Stecke and Solberg [1985] and Stecke [1986a]).

The rest of this paper 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 3.1, we prove pseudo-concavity of the throughput function for two special CQNs and make a conjecture that throughput is pseudo-concave for a general multi-server CQN. Section 3.2 states the Kuhn-Tucker necessary conditions and characteristics of solutions that satisfy these conditions. Two algorithms to find a workload allocation satisfying the necessary conditions are presented in Section 3.3. Computational experience of these algorithms is described in Section 4 for CQNs with numerous parameter values and the physical interpretation of the optimal workload is presented. 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 is considered here 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,..,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 $\bar{W} = (W_1,..,W_M)$, where W_i denotes the mean workload assigned to station i. The mean workload, W_i, is the mean service time demanded from station i by a customer in a typical cycle. This can be viewed as 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, demanded by a customer per visit there, 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 $\mu_i(j) = \min(j,S_i)$.

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

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

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

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

For this CQN, the throughput is then given by

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

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 (Buzen [1973]) or the mean value analysis (MVA) algorithm (Reiser and Lavenberg [1980]), 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:

P1:

Maximize $TH(N,\bar{W})$

subject to:

$$\sum_{i=1}^{M} W_i = TW, \qquad (3)$$

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

3. THE SOLUTION PROCEDURE

Let $\Gamma = \{\bar{W} \mid \Sigma_i W_i = TW;\ W_i \geq 0 \text{ for all } i\}$ denote the feasible region for Problem P1 and let $\Gamma' = \{\bar{W} \mid \Sigma_i W_i = TW;\ W_i > 0 \text{ for all } i\}$ denote the feasible interior region. The solution procedure is based upon the assumption that the throughput of a multi-server CQN is a pseudo-concave function of the workloads over Γ'. We show that this conjecture holds for two special cases: the single-server CQN, and the multi-server CQN for N=2.

3.1. Pseudo-concavity of Throughput

Stecke [1986b] shows that the throughput is not concave over Γ for a single-server CQN and conjectures that it is strictly quasiconcave. The basis for the conjecture is a proof of quasi-concavity for a single-server CQN with two stations and empirical evidence for a single-server CQN with three station. Stecke also provides some computational evidence that the function is strictly quasiconcave. We make a stronger conjecture that the function is pseudo-concave over Γ' for a multi-server CQN. (Pseudo-concavity implies that the function is strictly quasiconcave but the converse is not true.)

The benefit of assuming pseudo-concavity of the throughput function is that satisfaction of the first order conditions is both necessary and sufficient for optimality in the workload allocation problem. First we will show that the conjecture holds in some special cases.

Lemma 1. Let $g:C \rightarrow R^1$ and $h:C \rightarrow R^1$, where C 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 pseudo-concave. (3.41 on page 116 of Bazaraa and Shetty [1979])

Lemma 2. If function f is pseudo-concave, then $\bar{x}$ such that $\nabla f(\bar{x}) = 0$ is a global maximum of f. (page 106 of Bazaraa and Shetty [1979])

The property of Lemma 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 meeting the necessary conditions.

Lemma 3. TH(N,$\bar{W}$) is pseudo-concave over Γ' for a single-server CQN.

Proof: From the Little's Law, TH(N,$\bar{W}$) = N/C(N,$\bar{W}$) where C(N,$\bar{W}$) is the cycle time for the CQN. From Price [1974], we can easily show that C(N,$\bar{W}$) is a convex function of nonnegative $\bar{W}$ for a single-server CQN. Thus, it follows from Lemma 1 that TH(N,$\bar{W}$) is pseudo-concave over Γ'. ■

It is also clear that a CQN with only delay stations, that is, a CQN with $S_i \geq N$ for all i is pseudo-concave since TH(N,$\bar{W}$) = N/TW for any $\bar{W}$ for this CQN.

Lemma 4. TH(N,$\bar{W}$) is pseudo-concave over Γ' for a multi-server CQN when N=2.

Proof: We avoid considering two following cases from the previous discussion: (i) S_i=1 for all i and (ii) $S_i \geq 2$ for all i. Hence, without any loss of generality, we assume that S_i=1 for $1 \leq i \leq m$ and $S_i \geq 2$ for $m+1 \leq i \leq M$. Then, we can write TH(2,$\bar{W}$) as

$$\mathrm{TH}(2,\bar{W}) = \frac{\sum_{i=1}^{M} W_i}{\sum_{i\leq j} W_i W_j - \sum_{i=m+1}^{M} W_i^2/2} = \frac{TW}{TW^2/2 + \sum_{i=1}^{m} W_i^2/2} \cdot$$

The second equation is derived from the following substitutions: $\sum_{i=1}^{M} W_i = TW$ for $\bar{W} \in \Gamma'$ and $\sum_{i\leq j} W_i W_j = 1/2\,(\sum_{i=1}^{M} W_i)^2 + \sum_{i=1}^{M} W_i^2/2$. C(2,$\bar{W}$) = 2/TH(2,$\bar{W}$) is clearly convex over Γ' since each W_i^2 is convex and the sum of convex functions is convex. From Lemma 1, we prove this lemma. ■

These lemmas lead to the following conjecture.

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

As stated in Lemma 1, the sufficient condition for Conjecture 1 to hold true is to show that C(N,$\bar{W}$) is convex for a multi-server CQN. Empirical support for this conjecture is provided in the following sections.

3.2. Characteristics of Optimal Workloads

The Kuhn-Tucker necessary conditions for Problem P1 are given by

KT1:

$$\frac{\partial}{\partial W_i} TH(N,\bar{W}) + \nu + \pi_i = 0, \quad i=1,..,M,$$

$$\sum_{i=1}^{M} W_i = TW,$$

$$\bar{\pi} \geq \mathbf{0}, \ \bar{W} > \mathbf{0}, \ \bar{\pi} \cdot \bar{W} = 0,$$

where $\frac{\partial}{\partial W_i} TH(N,\bar{W})$ is the i^th^ element of the gradient vector of $TH(N,\bar{W})$ 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 $\frac{\partial}{\partial W_i} TH(N,\bar{W})$ can be expressed as (see, for example, Kobayashi and Gerla [1983]):

$$\frac{\partial}{\partial W_i} TH(N,\bar{W}) = -\frac{TH(N,\bar{W})}{W_i}(Q_i(N,\bar{W}) - Q_i(N-1,\bar{W})), \qquad (5)$$

where $Q_i(N,\bar{W})$ is the mean queue length at station i, including the customers in service, when there are N customers in the CQN.

Lemma 5. If $N \leq S_k = \max_i(S_i)$, then the optimum solution of Problem P1 is given by $\bar{W}^*$, where $W_k^* = TW$, and $W_j^* = 0$ for $j \neq k$. In this case, $TH(N,\bar{W}) = N/TW$.

Proof: From Little's law, the throughput is given as $TH(N,\bar{W}) = N/C(N,\bar{W})$, where $C(N,\bar{W})$ is the cycle time of the CQN. Clearly, $C(N,\bar{W}) \geq TW$. When $N \leq S_k$, station k is a delay station and no queueing takes places. Thus, the cycle time is TW when the total mean workload is assigned to station k. This obviously maximizes the throughput, and $TH(N,\bar{W}) = N/TW$ here. ■

If Conjecture 1 is true, then the solution described in Lemma 5 must satisfy the Kuhn-Tucker necessary conditions. This is stated as Lemma 6. A proof of Lemma 6 is provided in the Appendix.

Lemma 6. If $N \le S_k = \max_i(S_i)$, then $\bar{W}^*$ in Lemma 5 is the solution satisfying the necessary conditions given by KT1. ■

Having established that $\bar{W}^*$ in Lemma 5 satisfies the Kuhn-Tucker necessary conditions, we now show that it is a Brouwer's fixed point. This, in turn, leads to efficient solution procedures which are described in Section 3.3.

Definition 1. Let g be a continuous function such that $g:C \rightarrow C$, where $C \subset R^n$ is a convex and compact set. Then there exists an $\bar{x} \in C$ such that $g(\bar{x}) = \bar{x}$ by the Brouwer's Theorem (Todd [1976]). This $\bar{x}$ is called a Brouwer's fixed point.

In the following, we derive the form of $g(\bar{x})$. Suppose that $\bar{W} > \mathbf{0}$ satisfies the above necessary conditions for optimality. Then $\bar{\pi} = \mathbf{0}$, and $\bar{W}$, ν are the solution to the following equations:

$$\frac{\partial}{\partial W_i} TH(N,\bar{W}) + \nu \quad = \quad 0, \qquad i=1,..,M, \tag{6}$$

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

Multiplying both equations (5) and (6) by W_i, and summing over all i, we get

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

and so from equations (6) and (8),

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

Finally, from equations (5) and (9), we have

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

Letting $g_i(\bar{W}) = TW\,(Q_i(N,\bar{W}) - Q_i(N-1,\bar{W}))$, we have

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

where $g(\bar{W}) = (g_1(\bar{W}), ..., g_M(\bar{W}))$.

Lemma 7. The workload $\bar{W}$ satisfying equation (11) is a Brouwer's fixed point.

Proof: Let C be the feasible region Γ of Problem P1. Observe that C is an (M-1) dimensional simplex. Let the function g be defined as the mapping in equation (11). Now we will show that C and g satisfy Definition 1. Clearly C is a convex and compact set, and g is continuous over C, since $Q_i(N,\bar{W})$ is continuous over Γ for any nonnegative integer N. In order to show that $g:C \rightarrow C$, we will show that for any $\bar{W} \in C$, $g(\bar{W}) \in C$.

From results on the monotonicity of queue lengths (with respect to N) for a multi-server CQN (Suri [1984]), we have $g_i(\bar{W}) \geq 0$, for all i. Summing g_i over all i, we have

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

$$= TW(N-(N-1)) = TW.$$

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

If $\bar{W} > \mathbf{0}$ is a solution of equation (11), then it is a global optimum of Problem P1 under Conjecture 1, since it also satisfies the Kuhn-Tucker necessary conditions.

Lemma 8. Every extreme point of Γ is a Brouwer's fixed point satisfying equation (11).

Proof: Without loss of generality, choose an extreme point $\bar{W}$ such that $W_i = TW$, and $W_j = 0$, for all $j \neq i$. Then for each $j \neq i$, $g_j(\bar{W}) = 0$, since $Q_j(N,\bar{W}) = 0$ for any nonnegative N when $W_j=0$ and $g_i(\bar{W}) = TW - \sum_{j \neq i} g_j(\bar{W}) = TW - 0 = W_i$. Thus, this extreme point, $\bar{W}$, satisfies equation (11) and is a Brouwer's fixed point from Lemma 7. ■

Corollary 1. There are at least M Brouwer's fixed points satisfying equation (11) over the feasible region Γ.

Proof: The proof follows directly from Lemma 8. ■

However, every extreme point may not be a solution of KT1 as stated in Theorem 1. A proof of the theorem is given in the Appendix.

Theorem 1. If $N > \max_i (S_i)$, then any workload allocation $\bar{W}$ which has at least one $W_k=0$ cannot satisfy KT1, that is, the optimal workload allocation $\bar{W}^* > \mathbf{0}$. ■

The results of Lemmas 5 to 8 apply to a class of product form CQNs which are more general than the multi-server CQN. Lemmas 5 and 6 hold true for any product form CQN while Lemmas 7 and 8 hold for a CQN with N-monotonic stations (Suri [1984]). N-monotonic station i has the property that $P(n_i \geq k \mid n) \geq P(n_i \geq k \mid n-1)$ for all k and for $n \leq N$, where n_i denotes the number of customers in the system. It is clear that an N-monotonic station also has the property that the mean queue length is non-decreasing with n. Only the queue length monotonicity property is required to define a Brouwer's fixed point. It has been shown (Theorem 1 in Suri [1984]) that the multi-server CQN is a special case of the CQN with N-monotonic stations.

3.3. Algorithms for Problem P1

We coded two algorithms, the reduced gradient algorithm and the Eaves-Saigal fixed point algorithm, to solve Problem P1. Both algorithms use as an initial feasible point a balanced allocation (i.e., the total mean workload is allocated such that W_i/S_i is equal at all stations). 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 [1986a] 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 on two counts. 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, this algorithm requires the computation of M throughputs (to provide approximate partial derivative information) to determine which two W_i's are to be varied. The algorithm terminates when the sensitivity information indicates that further workload changes cannot increase the throughput.

The reduced gradient algorithms (Avriel [1976]) uses reduced gradient vectors by eliminating the dependent variables from the equality constraint, that is, equation (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 be changed after 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 P1 is equivalently rewritten as

P1':

$$\begin{array}{ll} \text{Minimize} & \theta(\bar{W}) \\ \text{subject to} & s(\bar{W}) \leq 0, \end{array}$$

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

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

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

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 KT1.

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

Proof: If $N > \max_i(S_i)$, then workload satisfying KT1, $\bar{W} > \mathbf{0}$ from Theorem 1. $\bar{W} > \mathbf{0}$ implies $s(\bar{W}) < 0$, that is, it is not a point on the boundary. Thus, $\mathbf{0} \in p(\bar{W})$ is equivalently stated as $\mathbf{0} = \nabla\theta(\bar{W})$. This workload $\bar{W}$ satisfies equation (11), and is a Brouwer's fixed point from Lemma 7. Therefore, the algorithm quadratically converges to it (Saigal [1977]). ■

The Eaves-Saigal algorithm appears to be the only algorithm with the property of quadratic convergence 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,..,M, each one of which is identical to CQN Ψ, but with station i removed.

4. 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 8, and the number of customers, N was specified as 5 or 20. Arbitrary unbalanced configurations for the server vector $\bar{S}$ were chosen. We chose each S_i to be less than N, since otherwise a trivial optimal solution is available from Lemma 5. The workloads were scaled such that $TW = \sum_{i=1}^{M} S_i$ without loss of generality (Stecke and Solberg [1985]).

We used the following measure for a termination condition:

$$D(\bar{W}) \;=\; \max_i | W_i - TW\,(Q_i(N,\bar{W}) - Q_i(N-1,\bar{W}))|. \qquad (13)$$

Clearly, $\bar{W} > 0$ satisfies the Kuhn-Tucker necessary conditions, KT1 when $D(\bar{W}) = 0$. The algorithms terminate when $D(\bar{W}) \leq \varepsilon$ for some specified tolerance ε. We specified five different tolerance levels in order to compare the speed of convergence of the two algorithms. They were 5.E-1, 1.E-1,1.E-2, 1.E-3, and 1.E-4. The following statistics were collected at each 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 termination. These statistics are summarized in Tables 1 and 2 below.

Tables 1 and 2.

For every problem in our experiment, both the algorithms always converged to the same interior point. We also tried other initial points but they still converged to the same interior point. Further, as we decreased the tolerance value for each problem, the throughput increased (see Tables 1 and 2) as expected. These observations lead to the following conjecture.

Conjecture 2. If $N > \max_i(S_i)$, then there is a unique solution, $\bar{W}$, for KT1.

The solution $\bar{W}$ is globally optimal under *either* Conjecture 1 *or* Conjecture 2 and it can be found by the Eaves-Saigal algorithm with quadratic convergence from Theorem 2. For small values of ε (i.e., $\varepsilon <$ 1.E-2), the Eaves-Saigal algorithm requires a far smaller number of throughput computations than the reduced gradient algorithm (see Table 1 and 2). Note that the reduced gradient algorithm was not executed for $\varepsilon <$ 1.E-2 due to its slow convergence. Also, our conjecture that the solution is unique is consistent with the observation made by Stecke and Solberg [1985] and Stecke [1986a] that there is a unique unbalanced optimal allocation for Problem P1.

Conjecture 2 and Lemma 8 lead to the following

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

Proof: The proof is given in the Appendix. ■

We mentioned earlier that $\bar{W} > 0$ satisfying equation (11) is the optimum solution for P1. Now we seek for an intuitive explanation of the equation. Equation (11) is rewritten as

$$Q_i(N,\bar{W}) - Q_i(N-1,\bar{W}) = \frac{W_i}{TW}, \quad i=1,..,M. \qquad (15)$$

When one customer is added to the system, the sum of the queue lengths among the stations in the system increases by one, and for the networks being considered, it can be shown (Suri [1984]) that the mean queue length at each station does not decrease. Equation (15) relates allocation of workloads among the stations to distribution of the queue length increase in the system at the N^{th} customer addition among M stations. Thus 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 workload at the corresponding station to the total workload.

In addition to the two algorithms, we also applied the method of successive substitution to equation (11). This method proceeds as follows: given an initial feasible point, it generates a sequence of points, all of which are guaranteed to be feasible (because of the form of (11) and the mean queue length monotonicity property). For all of the problems in our experiment, it always converged quickly to one of the extreme points. With different initial points, it converged to different extreme points. It appears that it converges to the nearest extreme point, following the steepest descent direction. Figures 1 and 2 graphically present this phenomenon for two-station CQNs. Both figures have only one interior point which is at the intersection of two curves, one plotting the function W_1/TW on the y-axis, and the other plotting the function $Q_1(N,W_1)$-$Q_1(N-1,W_1)$ on the y-axis. When the successive substitution method starts with a point to the left (right) of the intersection point, it quickly converges to the left (right) extreme point. The figures also show that the throughput is maximized at the point of intersection.

Figures 1 and 2.

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 P1, but the non-negativity constraints for W_i are replaced by constraints of the form

$$L_i \le W_i \le U_i,$$

where L_i and U_i are the lower and upper bounds, respectively, on W_i.

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. The following algorithm is used to find a good initial feasible solution. We assume that the indices of the stations are arranged so that $S_1 \geq S_2 \geq ... \geq S_M$. We also assume that there is a feasible workload allocation (i.e., $TW \leq \Sigma_i U_i$).

Algorithm 1.

1) Find a balanced workload allocation, $\bar{W}$. If it is feasible, then terminate. Otherwise go to step 2.

2) Let $E = \{ i \mid W_i > U_i \}$, $B = \{ i \mid W_i < L_i \}$, $S_E = \sum_{i \varepsilon E} (W_i - U_i)$, $S_B = \sum_{i \varepsilon B} (L_i. - W_i)$. Reset W_i to U_i for all $i \varepsilon E$ and to L_i for all $i \varepsilon B$. If $S_E - S_B > 0$ (less than the total workload is allocated), go to step 3. If $S_E - S_B < 0$ (more than the total workload, TW is allocated), go to step 4. Otherwise, terminate .

3) Reallocate $S_E - S_B$ by assigning as much additional workload as possible to stations 1,...,M in sequence while retaining feasibility. Terminate whenever a feasible reallocation has been found.

4) Reduce the workloads at stations M,...,1 in sequence while maintaining feasibility, until a total reduction of $S_B - S_E$ has been achieved.

The rationale for steps 3 and 4 is a result of Yao and Shanthikumar [1986] that for the multiple-server product-form CQN, throughput is increased by assigning more workload to a station with a larger number of servers.

Experimental Results

We conducted a number of experiments using the two nonlinear programming algorithms and Algorithm 1 described above, for CQNs with a range of parameter values. The value of M ranges from 2 to 8 and N is 5 or 20. We chose arbitrary unbalanced configurations for $\bar{S}$. Workloads were scaled such that the total 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.

Table 3.

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 $\nabla^2 \leq \varepsilon$ for some specified tolerance ε. We set ε = 1.E-5. Both algorithms were initialized with the solution from Algorithm 1. 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.

Table 4.

Algorithm 1 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 (7.b) with N=20. It appears that when the balanced workload allocation violates a bound, the optimal solution is to set the workload for the station equal to that bound. Thus, only the workloads of stations with inactive bound constraints need to be found using the NLP algorithm.

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 which has linear convergence. However, the Eaves-Saigal algorithm converges only linearly for problems with one or more bound constraints active at the optimum solution because of the manner in which constraints are handled by the algorithm.

6. CONCLUSION

In this paper, we considered the problem of allocating a given workload among the stations in a multi-server product form CQN to maximize the throughput. We first investigated the behavior of the throughput function and proved that it is pseudo-concave in two special cases: (i) the single-server CQN and (ii) the multi-server CQN when the number of customers is equal to 2. Computational experience with algorithms developed to find the optimal workload allocation also provided support for this conjecture in more general case. The advantage of having a pseudo-concave function is that the Kuhn-Tucker necessary conditions are sufficient for global optimality.

We also showed that if the number of customers in the CQN is greater than the number of servers in each station, the optimal workload is an interior Brouwer's fixed point and can be found by the Eaves-Saigal fixed point algorithm which has quadratic convergence (otherwise, we showed that it has a trivial optimum solution). Experimentally, we showed that this optimal workload is always unique. This 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 workload at the corresponding station to the total workload.

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.

REFERENCES

Avriel, M., 1976. *Nonlinear Programming Analysis and Methods*, Prentice-Hall, Inc.

Baskett, F., K.M. Chandy, R.R. Muntz and F.G. Palacios, 1975. Open, Closed, and Mixed Networks of Queues with Different Classes of Customers. *J. Assoc. Comput. Mach*, Vol 22, pp. 248-260.

Bazaraa, M. and C.M. Shetty, 1979. *Nonlinear Programming Theory and Algorithms*, John Wiley and Sons, Inc.

Buzen, J.P., 1973. Computational Algorithms for Closed Queueing Networks with Exponential Servers, *Commun. ACM*, Vol. 16, No. 9, pp. 527-531.

Gordon, W.J. and G.F. Newell, 1967. Closed Queueing Networks with Exponential Servers. *Operations Research*, Vol.15, pp. 252-267.

Kenevan, J.R. and A.K. Von Mayrhauser, 1984. Convexity and Concavity Properties of Analytic Queueing Models for Computer Systems, *Performance' 84*, E. Gelenbe (Editor), Elsevier Science Publishers B.V. (North-Holland) Amsterdam, pp. 361-375.

Kobayashi, H. and M. Gerla, 1983. Optimal Routing in Closed Queueing Networks, *ACM Transactions on Computer Systems*, Vol. 1, No. 4, pp. 294-310.

Morin, T.L. and K.E. Stecke, 1985. The Optimality of Balancing Workloads in Certain Types of Flexible Manufacturing Systems, *European J. of Operational Research*, Vol. 20, pp. 68-82.

Price, T.G., 1974. Probability Models of Multiprogrammed Computer Systems. Ph.D. dissertation, Department of Electrical Engineering, Stanford University, Stanford CA.

Reiser, M. and S. Lavenberg, 1980. Mean-Value Analysis of Closed Multichain Queueing Networks, *Journal of the Association for Computing Machinery*, Vol. 27, No. 2, pp. 313-322.

Saigal, R., 1977. On the Convergence Rate of Algorithms for Solving Equations That Are Based on Methods of Complementary Pivoting. *Math. Oper. Res.*, Vol. 2, pp. 108-24.

Secco-Suardo, G., 1978. Optimization of a Closed Network of Queues, Report No. ESL-FR-834-3, Electronic Systems Laboratory, M.I.T., Cambridge,

Shanthikumar, J.G. and D.D. Yao, 1987. Optimal Server Allocation in a Multi-server Stations, *Management Science*, Vol. 33, No. 9, pp. 1173-1180.

Solberg, J.J., 1979. Stochastic Modeling of Large Scale Transportation Networks, Report No. DOT-ATC-79-2, School of Industrial Engineering, Purdue University, West Lafayette IN.

Stecke, K.E., 1983. Formulation and Solution of Nonlinear Integer Production Planning Problems for Flexible Manufacturing Systems, *Management Science*, Vol. 29, No. 3, pp. 273-288.

Stecke, K.E., 1986a. A Hierarchical Approach to Solving Machine Grouping and Loading Problems of Flexible Manufacturing Systems, *European J. of Operational Research*, Vol. 24, pp. 369-378.

Stecke, K.E., 1986b. On the Nonconcavity of Throughput in Certain Closed Queueing Networks, *Performance Evaluation*, Vol. 6, No. 4, pp. 293-305.

Stecke, K.E. and J.J. Solberg, 1985. The Optimality of Unbalancing Both Workloads and Machine Group Sizes in Closed Queueing Networks of Multi-Server Queues, *Operations Research*, Vol. 33, No. 4, pp. 882-910.

Suri, R., 1984. A Concept of Monotonicity and Its Characterization for Closed Queueing Networks, *Operations Research*, pp. 606-624.

Todd, M.J., 1976. *The Computation of Fixed Points and Applications*, Springer-Verlag, Berlin-Heidelberg.

Trivedi, K.S. and R.E. Kinicki, 1978. A Mathematical Model for Computer System Configuration Planning, *Performance of Computer Installations*, D. Ferrari (Editor), North-Holland, Amsterdam.

Trivedi, K.S. and T.M. Sigmon, 1981. Optimal Design of Linear Storage Hierarchies, *Journal of the Association for Computing Machinery*, Vol. 28, No. 2, pp. 270-288.

Trivedi, K.S. and R.A. Wagner, 1979. A Decision Model for Closed Queueing Networks, *IEEE Transactions on Software Engineering*, Vol. 5, No. 4, pp. 328-332.

Trivedi, K.S., R.A. Wagner and T.M. Sigmon, 1980. Optimal Selection of CPU Speed, Device Capabilities and File Assignments, *Journal of the Association for Computing Machinery*, Vol. 27, No. 3, pp. 457-473.

Yao, D.D., 1985. Some Properties of the Throughput Function of Closed Networks of Queues, *Operations Research Letters*, Vol. 3, No. 6, pp. 313-317.

Yao, D.D. and J.G. Shanthikumar, 1986. On Server Allocation in Multiple Center Manufacturing Systems. Dept. of Industrial Engineering and Operations Research, Columbia University, New York.

APPENDIX

Lemma 6. If $N \leq S_k = \max_i (S_i)$, then $\bar{W}$ such that W_k=TW and W_j=0 for all $j \neq k$, is the solution of the necessary conditions, KT1.

Proof: From equation (5) and $\lim_{W_k \to TW} Q_k(N,\bar{W}) = N$, we have

$$\lim_{W_k \to TW} \frac{\partial TH(\bar{W})}{\partial W_k} = -\frac{TH(\bar{W})}{TW}(N-(N-1)) = -\frac{TH(\bar{W})}{TW} = -\frac{N}{TW^2}.$$

From Little's law, $Q_j(N,\bar{W}) = v_j TH(N,\bar{W}) R_j(N,\bar{W})$ where v_j is the visit frequency at station j and $R_j(N,\bar{W})$ is mean queueing time at station j including service time, τ_j. As $W_j \to 0$ for $j \neq k$, $R_j(N,\bar{W}) \to \tau_j$ since with very small workload assigned to station j, an arriving customer very likely finds an available server among S_j and stays at station j during service time only. Thus we have

$$\lim_{W_j \to 0} \frac{\partial TH(\bar{W})}{\partial W_j} = \lim_{W_j \to 0} -\frac{TH(N,\bar{W})}{W_j}(v_j TH(N,\bar{W})R_j(N,\bar{W}) - v_j TH(N-1,\bar{W})R_j(N-1,\bar{W}))$$

$$= \lim_{W_j \to 0} -\frac{TH(N,\bar{W})}{W_j}(v_j TH(N,\bar{W})\tau_j - v_j TH(N-1,\bar{W})\tau_j)$$

$$= - TH(N,\bar{W})\,(\,TH(N,\bar{W}) - TH(N-1,\bar{W})\,) \quad \text{from } W_j = v_j \tau_j$$

$$= -\frac{N}{TW}\left(\frac{N}{TW} - \frac{N-1}{TW}\right) \quad \text{from Lemma 5}$$

$$= -\frac{N}{TW^2}.$$

Given the derivative values at $\bar{W}$, we now solve the necessary conditions, KT1. $W_k > 0$ implies π_k=0. Thus ν is given as $\nu = -\frac{\partial TH(N,\bar{W})}{\partial W_k} = \frac{N}{TW^2}$ and $\pi_j = -\frac{\partial TH(N,\bar{W})}{\partial W_j} - \nu = 0$ for each $j \neq k$. Therefore, $\bar{W}$, $\bar{\pi}$, and ν obtained above is the solution of KT1. ■

Lemma 9. $C(N+1,\bar{W}) \geq C(N,\bar{W})$ for a multi-server CQN.

Proof: $C(N,\bar{W})$, cycle time of the CQN with N customers is written as $C(N,\bar{W}) = \sum_{i=1}^{M} R_i(N,\bar{W})$ where $R_i(N,\bar{W})$ is mean queueing time at station i including service time, τ_i. We prove this lemma by showing that $R_i(N+1,\bar{W}) \geq R_i(N,\bar{W})$ for all i. From equation. (2.19) of Reiser and Lavenberg [1980],

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

where $p_i(n|N)$ is probability that n customers are present at station i for the CQN with N customers. 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

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

Denoting $\sum_{n=S_i}^{N} p_i(n|N)$ as $p_i(n \geq S_i \mid N)$, we rewrite $\Delta R_i(N,\bar{W})$ as

$$\Delta R_i(N,\bar{W}) = \frac{\tau_i}{S_i}[(\,p_i(n \geq S_i \mid N) - p_i(n \geq S_i \mid N-1)\,) + \sum_{k=S_i+1}^{N} (\,p_i(n \geq k \mid N) - p_i(n \geq k \mid N-1)\,)]$$

$$= \frac{\tau_i}{S_i}\sum_{k=S_i}^{N} (\,p_i(n \geq k \mid N) - p_i(n \geq k \mid N-1)\,),$$

where $p_i(n \geq k \mid N) - p_i(n \geq k \mid N-1) \geq 0$ for all k since a multi-server CQN is a special class of CQNs which have all stations ∞-monotonic (Suri [1984]). Therefore,we have shown $\Delta R_i(N,\bar{W}) \geq 0$ for all i and we prove the lemma. ■

Theorem 1. If $N > \max_i (S_i)$, then any workload allocation $\bar{W}$ which has at least one $W_k=0$ cannot satisfy KT1, that is, the optimal workload allocation $\bar{W}^* > \mathbf{0}$.

Proof: Suppose that $\bar{W}$ which has $W_k=0$ for some k satisfies KT1. Let $I=\{i \mid W_i > 0\}$ and $\bar{I}=\{k \mid W_k = 0\}$. Clearly, $\bar{I}$ is not empty. From KT1, for each i in I, $\pi_i = 0$ and

$$TH(N,\bar{W})(\,Q_i(N,\bar{W}) - Q_i(N-1,\bar{W})\,) = \nu\, W_i \qquad \text{for } i \,\varepsilon\, I.$$

Summing the above equation over all i in I, we have

$$TH(N,\bar{W})\ (\ N - (N-1)\ ) = TH(N,\bar{W}) = \nu \sum_{i\, \varepsilon I} W_i = \nu\ TW.$$

That is, $\nu = TH(N,\bar{W})/TW$. As shown in Lemma 6, we have

$$\lim_{W_k \to 0} \frac{\partial TH(N,\bar{W})}{\partial W_k} = -\ TH(N,\bar{W})\ (\ TH(N,\bar{W}) - TH(N-1,\bar{W})\ ) \text{ for each } k\ \varepsilon\ \bar{I}.$$

Subsequently, we have from KT1 and substitution of ν

$$\begin{aligned} \pi_k &= -\ \nu + TH(N,\bar{W})\ (\ TH(N,\bar{W}) - TH(N-1,\bar{W})\ ) \\ &= TH(N,\bar{W})\ \{\ (\ TH(N,\bar{W}) - TH(N-1,\bar{W})\ ) - 1/TW\ \} \text{ for each } k\ \varepsilon\ \bar{I}. \end{aligned}$$

We now show that $\pi_k < 0$ for each $k\ \varepsilon\ \bar{I}$, so that $\bar{W}$ does not satisfy KT1. Denote $C(N,\bar{W})$ as the cycle time of the CQN with N customers. When $N > \max_k (S_k)$, $C(N,\bar{W}) > TW$ since there is positive probability that all customers are queued at one station and some customers have waiting time before being served.

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

which is < 0 since $1/C(N,\bar{W}) - 1/TW < 0$, and $1/C(N,\bar{W}) - 1/C(N-1,\bar{W}) \leq 0$ from Lemma 9. Thus, we have shown that $\pi_k < 0$ for each $k\ \varepsilon\ \bar{I}$ and we prove the lemma. ■

Proposition 1. If $N > \max_i (S_i)$, then there are exactly 2^M -1 Brouwer's fixed points satisfying equation (11) over the feasible region of Problem P1 when Conjecture 2 holds.

Proof: Firstly, we know from Lemma 8 that there are M Brouwer's fixed points where one station has all the workload assigned to it. Now, consider loading only the first two stations in the CQN and fixing W_i to 0 for i=3 to M. Solve Problem P1 for this two station sub-CQN. Since $N > \max (S_1,S_2)$, there is only one solution (W_1,W_2) of the KT1 for the reduced problem from Conjecture 2, which is an interior point such that $W_1 > 0$, $W_2 > 0$ and $W_1+W_2 = TW$. Clearly, $\bar{W} = (W_1,W_2,0,...,0)$ is a Brouwer's fixed point satisfying equation (11) for the original problem. There are $\binom{M}{2}$ such Brouwer's fixed points by choosing two stations out of M. Applying a similar argument, we can show that there are $\binom{M}{p}$ Brouwer's fixed points satisfying equation (11) for the original problem in which there are p components having a positive value for W_i. Therefore, the total number of Brouwer's fixed points satisfying equation (11) is given as

$$\sum_{p=1}^{M} \binom{M}{p} = 2^M - \binom{M}{0} = 2^M - 1. \qquad \blacksquare$$

Table 1. Reduced Gradient Algorithm vs. Eaves-Saigal Algorithm at N=5

System Config.	Terms	$D(\bar{W}) = \max_i \lvert W_i - TW(Q_i(N,\bar{W}) - Q_i(N-1,\bar{W}))\rvert$						
		Reduced Gradient			Eaves-Saigal			
		5.E-1	1.E-1	1.E-2	1.E-1	1.E-2	1.E-3	1.E-4 or less
M=2	th§			.8421779	.8421582			.8421872
$\bar{S}$=	no†			9	4			8
(1,3)	∇^2‡			1.E-4	4.E-5			2.E-12
M=3	th	.6462935	.6537918	.6539210		.6539216		.6539243
$\bar{S}$=	no	24	36	58		10		15
(1,2,4)	∇^2	5.E-4	3.E-5	4.E-7		1.E-6		3.E-13
M=4	th	.4760777	.4803953	.4805868			.4805915	.4805916
$\bar{S}$=	no	11	19	21			14	20
(2,2,2,4)	∇^2	5.E-5	3.E-6	3.E-7			8.E-10	2.E-13
M=5	th	.3480955	.3538133	.3541686		.3541608		.3541713
$\bar{S}$=	no	5	113	233		16		48
(1,3,3,3,4)	∇^2	8.E-4	5.E-7	2.E-7		1.E-6		3.E-13
M=6	th	.3033259	.3098448	.3109342		.3109373		.3109450
$\bar{S}$=	no	3	73	243		60		76
(1,2,2,3,4,4)	∇^2	5.E-4	4.E-6	4.E-7		8.E-7		3.E-11
M=7	th	.2708771	.2757104	.2760962			.2760995	.2761004
$\bar{S}$=	no	1	165	352			41	59
(1,2,2,3,3,3,4)	∇^2	3.E-4	1.E-6	3.E-7			7.E-8	5.E-12
M=8	th	.4128372	.4154144	.4155468		.4155474	.4155482	
$\bar{S}$=	no	119	151	226		42	52	
(1,1,1,1,1,1,1,4)	∇^2	6.E-5	3.E-6	2.E-7		9.E-8	4.E-9	

§ th = throughput
† no = the number of throughput computations
‡ ∇^2 = the second norm of the steepest ascent feasible direction

Table 2. Reduced Gradient Algorithm vs. Eaves-Saigal Algorithm at N = 20

System Config.	Terms	$D(\bar{W}) = \max_i \lvert W_i - TW(Q_i(N, \bar{W}) - Q_i(N-1, \bar{W}))\rvert$							
		Reduced Gradient			Eaves-Saigal				
		5.E-1	1.E-1	1.E-2	5.E-1	1.E-1	1.E-2	1.E-3	1.E-4 or less
M = 2	th§			.9599658	.9596465				.9599665
$\bar{S}$=	no†			9	6				10
(1,3)	∇^2‡			5.E-6	2.E-3				3.E-10
M = 3	th	.9130309	.9137392	.9137404	.9134154	.9136574	.9137410	.9137411	.9137412
$\bar{S}$=	no	16	22	30	4	9	19	24	29
(1,2,4)	∇^2	7.E-4	9.E-6	3.E-6	1.E-3	4.E-4	8.E-8	5.E-9	1.E-12
M = 4	th	.8558376	.8559553	.8559903		.8559714		.8559908	.8559908
$\bar{S}$=	no	12	20	56		5		13	19
(2,2,2,4)	∇^2	3.E-4	7.E-5	9.E-7		9.E-6		3.E-9	1.E-11
M = 5	th	.7983813	.7984879	.7985132		.7984711	.7985132		.7985133
$\bar{S}$=	no	34	47	151		10	25		40
(1,2,3,4, 7)	∇^2	8.E-5	3.E-5	9.E-8		5.E-5	3.E-8		1.E-12
M = 6	th	.7341363	.7342175	.7342763		.7342627	.7342759		.7342773
$\bar{S}$=	no	48	72	232		24	32		49
(1,2,2,3, 6,8)	∇^2	2.E-5	5.E-5	1.E-8		9.E-6	9.E-7		1.E-9
M = 7	th	.7225042	.7229673	.7229980	.7229260		.7229979		.7229986
$\bar{S}$=	no	20	50	179	14		25		46
(1,2,2,3, 3,3,4)	∇^2	5.E-5	1.E-5	8.E-8	5.E-5		5.E-8		1.E-12
M = 8	th		.6592568	.6592672		.6592500	.6592676	.6592687	
$\bar{S}$=	no		77	274		50	60	81	
(1,1,2,2, 3,3,5,9)	∇^2		4.E-6	5.E-9		5.E-6	6.E-8	1.E-9	

§ th = throughput
† no = the number of throughput computations
‡ ∇^2 = the second norm of the steepest ascent feasible direction

Figure 1. Throughput and Marginal Queue Length Increase for the CQN with S=(1,3) and N=10

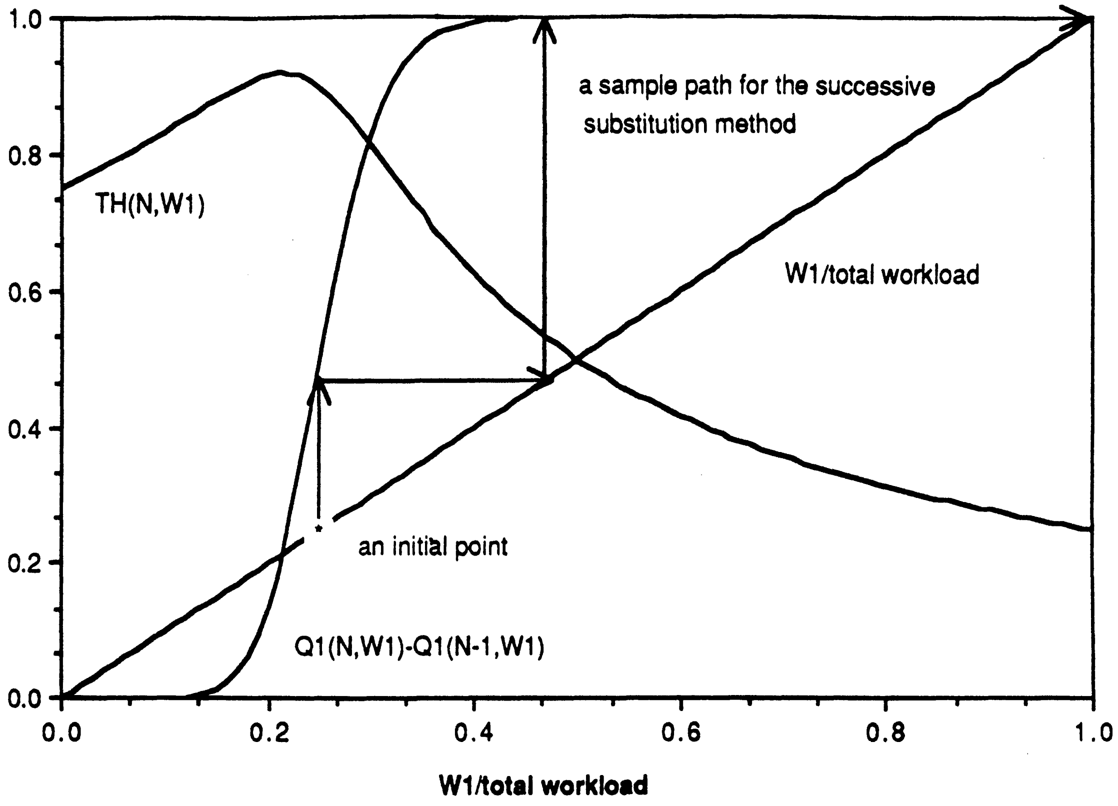

Figure 2. Throughput and Marginal Queue Length Increase for the CQN with S=(5,2) and N=10

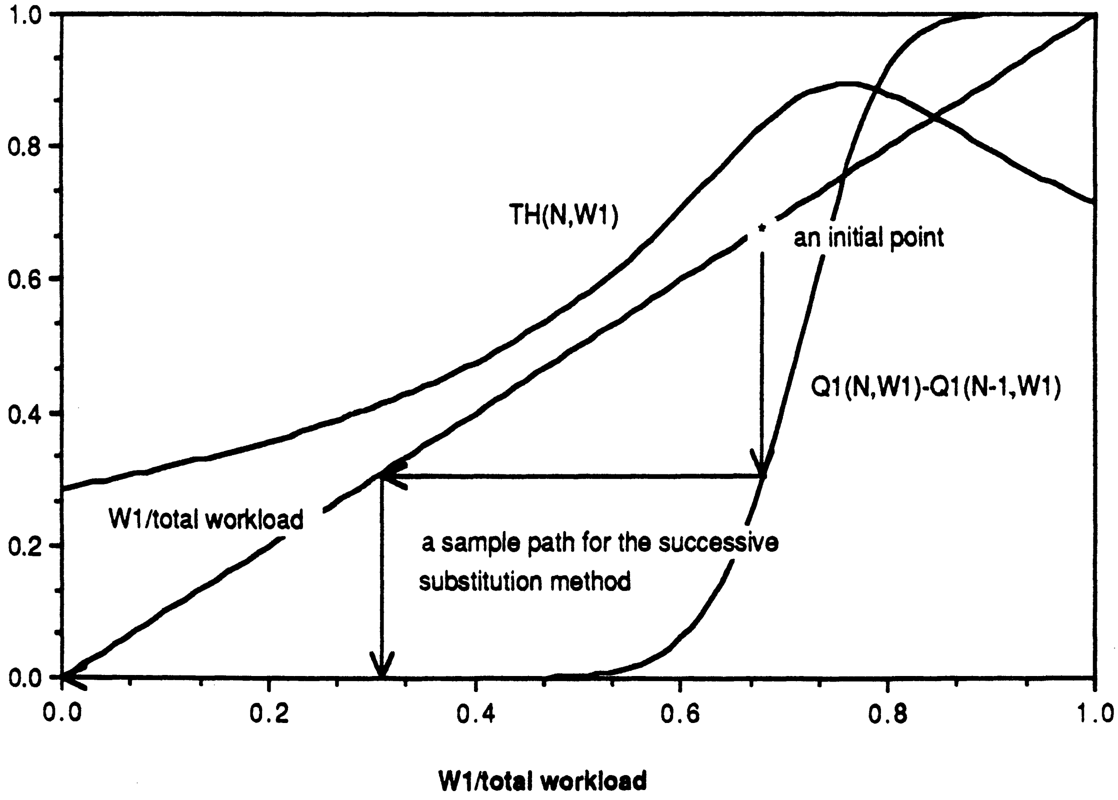

Table 3. Seven Sets of Problem P1 with Loose (a) and Tight (b) Bound Constraints

(1) M=2, $\bar{S}$=(3,1)

a) $\bar{L}$=(.1,.1), $\bar{U}$=(4.,3.)

b) $\bar{L}$=(2.,1.), $\bar{U}$=(4.,3.)

(2) M=3, $\bar{S}$=(4,2,1)

a) $\bar{L}$=(.5,.5,.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.,.1,.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) $\bar{L}$=(.9,.8,.7,.6.,.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.,.9,.8,.7,.6,.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) $\bar{L}$=(.5,.5,.5,.5,.5,.5,.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) $\bar{L}$=(.5,.5,.5,.5,.5,.1,.1,.1), $\bar{U}$=(21.,21.,21.,21.,21.,21.,21.,21.)

b) $\bar{L}$=(2.,2.,2.,2.,3.,1.,.5,.5), $\bar{U}$=(20.,20.,20.,20.,20.,20.,20.,20.)

Table 4. Reduced Gradient Algorithm vs. Eaves-Saigal Algorithm

Problem	Terms	Reduced Gradient Algorithm		Evase-Saigal Algorithm	
		N = 5	N = 20	N = 5	N = 20
(1.a)	th*	.8421862	.9599664	.8421872	.9599664
	no§, act#	31, 0	6, 0	6, 0	8, 0
(1.b)	th	.7954545	.9497207	.7954545	.9497207
	no, act	1, 1	1, 1	1, 1	1, 1
(2.a)	th	.6509428	.9137390	.6508728	.9137412
	no, act	28, 0	44, 0	10, 0	10, 0
(2.b)	th	.5457154	.6663790	.5457154	.6663790
	no, act	1, 2	1, 2	1, 2	1, 2
(3.a)	th	.4798654	.8559814	.4803098	.8559714
	no, act	61, 0	18, 0	11, 0	6, 0
(3.b)	th	.4661922	.8492100	.4661922	.8492100
	no, act	1, 3	1, 3	1, 3	1, 3
(4.a)	th	.2913062	.7984549	.2925165	.7985078
	no, act	8, 1	64, 0	36, 0	19, 0
(4.b)	th	.2722947	.4994284	.2722947	.4994284
	no, act	1, 3	1, 3	1, 3	1, 3
(5.a)	th	.2255052	.7340815	.2265839	.7342201
	no, act	2, 1	67, 0	53, 0	33, 0
(5.b)	th	.2229083	.6201884	.2229083	.6201884
	no, act	1, 4	1, 4	1, 4	1, 4
(6.a)	th	.2738760	.7229273	.2757980	.7229711
	no, act	2, 1	48, 0	57, 0	16, 0
(6.b)	th	.2588161	.4981604	.2588161	.4981610
	no, act	1, 5	1, 5	1, 5	1, 5
(7.a)	th	.1913671	.6587587	.1908631	.6592500
	no, act	3, 1	53, 0	4, 0	51, 0
(7.b)	th	.1904921	.5984791	.1904921	.5992992
	no, act	1, 3	42, 1	1, 3	80, 0

* th = throughput

§ no = the number of throughput computations

act = the number of active bound constraints at the termination point