Division of Research
Graduate School of Business Administration
The University of Michigan

November 1982

Revised, February 1984

THE OPTIMALITY OF UNBALANCING BOTH WORKLOADS AND MACHINE GROUP SIZES IN CLOSED QUEUEING NETWORKS OF MULTI-SERVER QUEUES

Working Paper No. 322-b

Kathryn Stecke
The University of Michigan

James J. Solberg
Purdue University

Operations Research, forthcoming, 1985

ABSTRACT

A closed queueing network model is used to explore the consequences of varying workloads among multi-server queues, not necessarily of equal sizes. In addition, the problem of assigning servers of similar types to the queues in the network, to maximize expected throughput, is solved.

It is shown that: (1) unbalanced configurations of assigned servers are superior to balanced; and (2) unbalanced workloads are better than balanced. Differences in system throughput from balanced versus unbalanced configurations/workloads can be significant. Applications to planning problems of flexible manufacturing systems are discussed.

Subject classification: 694 and 696-optimization of multi-server closed queueing networks. 581-planning and control of flexible manufacturing systems.

There is an interesting conflict in firmly held beliefs about the desirability of balancing workloads in production systems. On the one hand, some authors have considered it so self-evident that an assignment which equalizes the workload per machine is optimal, that that condition is taken as the objective in an optimization problem, and all attention is focused on the constraints which might interfere with attaining the goal. The large body of literature on assembly line balancing (see, e.g., Ignall [1965]) adopts this point of view. On the other hand, it is well known among queueing theorists that, under stochastic conditions, a pooled group of servers will perform more efficiently (i.e., complete more jobs per unit time or increase production rate) than the same number of servers working separately (see, e.g., Kleinrock [1976]). This fact suggests that an optimal assignment might be to provide proportionately greater workloads to the servers of a large group than to the servers of a smaller group or to an individual server. This paper explores the issue in the context of a closed queueing network model which has proved useful in representing a wide variety of real manufacturing systems.

There are actually two questions to be answered. The first is: Given the freedom to arrange the available servers into groups of varying sizes, is it optimal to equalize the number of servers in each group? The second is: Given a specified configuration, which may or may not involve equal-sized groups, is it optimal to equalize the workload per server? As we shall see, the answers indicate that balancing is hardly ever optimal.

In particular, the problem of how to best partition m servers into g groups has not to our knowledge previously been investigated. This problem, which we call the grouping problem, is addressed here. The solution is that all possible partitions of m servers into g groups can be ordered according to the maximum expected production of the system. In particular, the solution to the grouping problem is as follows:

i) The more pooling (or the fewer groups), the better;

ii) If constraints indicate that g groups of m servers are required, then the more unbalanced configuration provides a larger maximum expected production.

In addition, we address the problem of optimally allocating a total amount of work among a system of grouped servers, and call this the loading problem. The solution to the loading problem is:

i) For a system comprised of groups of unequal sizes (the better configurations), the expected production rate is maximized by assigning a specific unbalanced workload per server to each group;

ii) Balancing the workload per server is optimal only if all group sizes are equal.

Several previous studies of stochastic production lines have pointed out that balancing is nonoptimal in serial systems of single-server queues with finite buffers (Makino [1964], Hillier and Boling [1966, 1967, 1979], Rao [1976], Magazine and Silver [1978], El-Rayeh [1979]). However, this phenomenon is related to the finite buffer condition, rather than to the multiple-server efficiency issue which is studied here. In manufacturing terms, this previous work related to assembly lines, whereas our results would pertain to job shops and flexible manufacturing systems (FMSs). Indications that highly unbalanced workload assignments could be significantly better than the balanced solutions came in an empirical study of an actual FMS (Stecke [1977], Stecke and Solberg [1981b]).

The closed queueing network (CQN) model which forms the basis of our theoretical study is a product form network. Its value and accuracy in modeling real manufacturing systems was revealed in Solberg (1977) and confirmed in several subsequent investigations (Secco-Suardo [1978], Shanthikumar and Buzacott [1979], Hildebrandt [1980], Buzacott and Shanthikumar [1980], Cavaillé and Dubois [1982]).

The paper is organized as follows. The closed queueing network model, its robustness, and our application of the model are presented in the following section. Results concerning the optimality of maximum pooling are given in §2. §3 contains results associated with optimum partial pooling as well as workload allocation in a partially pooled system. The results and their applicability to flexible manufacturing are discussed in §4. The final section suggests directions for future research.

1. THE CLOSED QUEUEING NETWORK MODEL

The description of the CQN model as well as its robustness are given below. Manufacturing terminology is used, since that is our main application.

1.1 Definitions and Notation

One depiction of a CQN as a system of arbitrarily-connected machine groups is given in Figure 1. There are g groups in the system and s_i machines in group i. The average processing time of an operation by one of the machines in group i is t_i. Note that $m = \sum_{i=1}^{g} s_i$, since we seek to partition m machines into g groups. For flexible manufacturing applications, transporters, as well as loading and unloading (L/UL) stations, can be considered by using the central-server model, as shown in Figure 2, which is a special case of our CQN model. However, the remainder does not include the transporter and L/UL in order to solve the theoretical, continuous loading problem addressed here: the determination of the optimal allocation of a total amount of work among a system of grouped machines. In effect, the analysis and calculations which follow are as if the transporter and L/UL operations were infinitely fast.

Routing through the system is arbitrary. The routing could be first-order Markovian (defined by transition probabilities, p_{ij}--see Figure 1); multiple-class Markovian (defined by transition probabilities for each part type k,

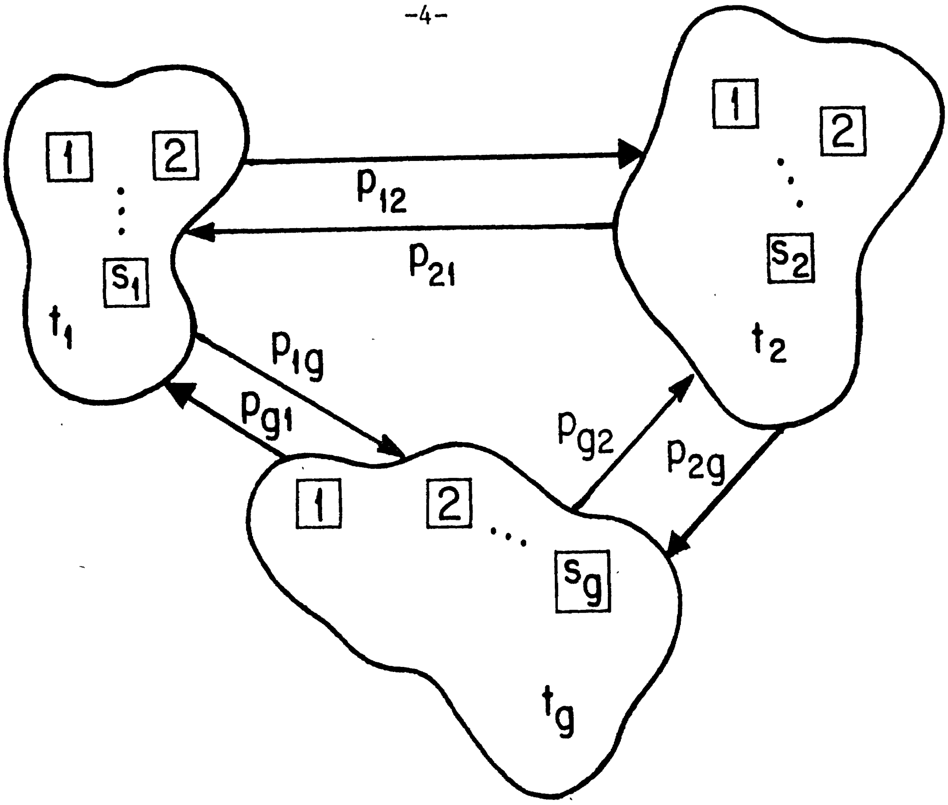

Figure 1. An Arbitrarily-Connected Network of Multi-Server Queues.

$p_{ij}(k)$ (Reiser and Kobayashi [1975])); higher-order Markovian (for example, second order is defined by p_{ijk}, which is the probability that a part previously at i, now at j, goes next to k (Kobayashi and Reiser [1975])); and even fixed routes through the system (defined by routing vectors for each part type, r(k) = (r(k,1), r(k,2),...), where r(k,j) is the index of the j'th machine group visited by a part of type k (Kelly [1979])). All these routing mechanisms produce the same values for certain output measures as do the q_i, which are shown in Figure 2. For additional routing information, see Stecke and Solberg (1981a).

In particular, the q_i's in Figure 2 are relative arrival rates to the various machine groups, i.e., any non-negative solution to the traffic equations, $q_i = \sum_j p_{ji} q_j$, where the p_{ji}'s are routing probabilities. Our formulas permit any scaling of the q_i's. In particular, if the q_i's are scaled to sum to one, q_i can be interpreted as both the probability that the next machine visited is one in group i and the average number of visits to a machine

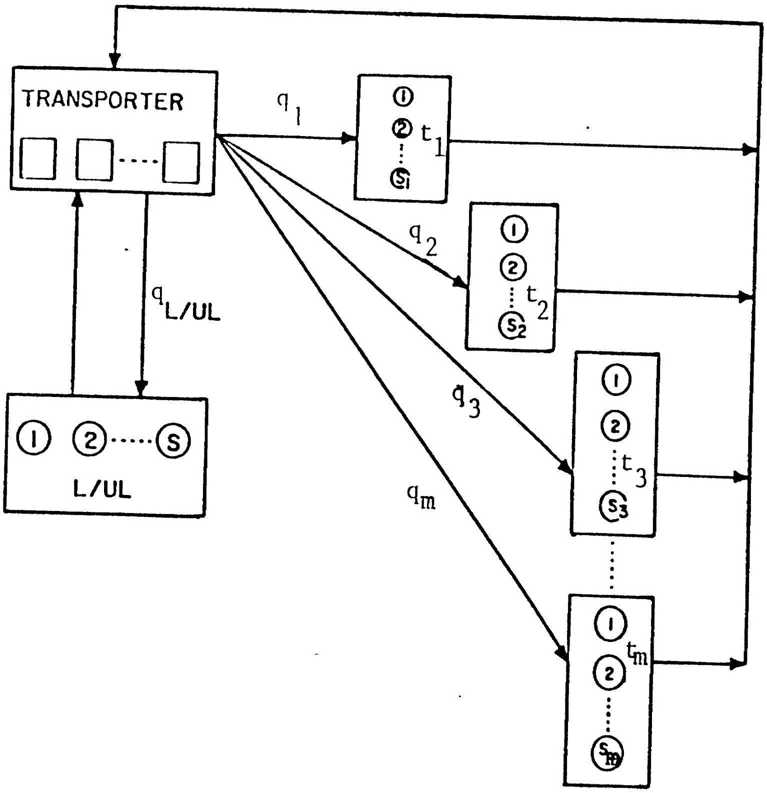

Figure 2. Central-Server Model.

in group i per visit to a transporter. For a different scaling, q_i is the mean number of operations per part that visit machine group i.

Since the system is closed, it always contains a fixed number of parts, n. When a part is finished it can be thought of as being taken to a load/unload station, where another part immediately enters the system.

The queue discipline can be either FCFS, infinite server, LCFS preempt-resume, processor sharing (see Baskett et al. [1975]), random selection (see Spirn [1979]), or one developed by Kelly (1979) that allows an arbitrary distribution to be defined at each node. The service time distribution is arbitrary, except for FCFS machine groups, which require exponential service times. However, robustness is observed in the FCFS case even when service times are not exponentially distributed (see Suri [1983]).

In fact, queueing network models have been found to be surprisingly accurate in predicting steady-state behavior of flexible manufacturing systems despite requiring only average processing times and visit frequencies to define workload parameters for each group. One might expect that more information would be required. However, Solberg (1977) compared results from his model, CAN-Q, to those of a very detailed simulation of a particular FMS, the Sundstrand/Caterpillar system (Stecke [1977]), and found that output measures of average machine utilizations and production rate all differed by less than 3 percent. The network-of-queues model has also been used, with good results, to model other nonmanufacturing systems where the service time distribution was not exponential. Hughes and Moe (1973), Buzen (1975), Giammo (1976), Lipsky and Church (1977), Rose (1976, 1978), and Suri (1983) have validated queueing network models through empirical studies, and have verified that the models reproduced observed quantities with great accuracy.

1.2 Our Variant of the CQN Model

For our purposes, it suffices to restrict attention to single-class closed queueing networks consisting of a single type of identical or similar parts. This is sufficient since our theoretical loading problem is to allocate a total amount of work among a system of machine groups. Multi-class models can be used, at additional computational expense, to handle multiple part types.

The usual measure of the relative workload assigned to group i is w_i (Solberg [1977], Buzen [1971, 1973], Reiser and Kobayashi [1975]), where $w_i = q_i t_i$, $i = 1,\ldots,g$. These workloads are relative, since the q_i's need not sum to one. For our purposes, w_i was scaled to provide our workload measure:

$$X_i = q_i t_i / [(\sum_{j=1}^{g} q_j t_j)/(\sum_{j=1}^{g} s_j)] \quad . \tag{1}$$

Notice that the numerator is the usual definition of workload assigned to group i, and the denominator is the average workload per machine.

This scaling was chosen for several reasons:

i) For a given number of machines in the system, regardless of their grouping, the total amount of work to be allocated among groups always equals the total number of machines:

$$\sum_{i=1}^{g} X_i = \sum_{i=1}^{g} s_i = m;$$

ii) The workload is independent of any chosen scaling of q_i;

iii) We compare alternative workloads to a balanced workload. For a balanced workload, regardless of system size or configuration,

$$X_1/s_1 = X_2/s_2 = \ldots = X_g/s_g = 1.$$

X_i/s_i is interpreted as being proportional to the utilization (fraction of time busy) of a typical machine in group i. Balanced means that all individual machines have identical utilizations;

iv) This scaling provides a normalized, dimensionless evaluator, a production function (the subsequently defined Pr(g,n;S,X) of equation (4)), whose values lie between zero and one, while X_i lies between zero and m. A value of one corresponds to a perfect system, when all machines are busy all the time, a situation which would never happen in practice. This situation "occurs" in a perfectly balanced system with an infinite number of parts in the system.

The workload definition (1) allows new alternative, equivalent definitions of the production function, which are useful in proving properties of the function. These are defined and described in Stecke (1981) and Stecke and Schmeiser (1983). Finally, the workload, X_i, can be allocated differently either by fixing q_i and varying t_i, or vice versa.

The state of the system is given by $\tilde{n}$, which is $(n_1,n_2,\ldots,n_g)$, where n_i is the number of parts at machine group i, both those waiting and those in process. For all i, n_i is an integer between zero and n, and $\sum_{i=1}^{g} n_i = n$. The steady-state probability of being in state $\tilde{n}$ is $p(\tilde{n}) = p(n_1,n_2,\ldots,n_g)$, which has the product form solution: $p(\tilde{n}) = \frac{1}{G(g,n;S,X)} \prod_{i=1}^{g} f_i(n_i)$, where $S = (s_1,s_2,\ldots,s_g)$, $X = (X_1,X_2,\ldots,X_g)$, the normalizing constant is:

$$G(g,n;S,X) = \sum_{\substack{n_1 \geq 0 \quad n_2 \geq 0 \ \ldots \ n_g \geq 0 \\ n_1 + n_2 + \ldots + n_g = n}} f_1(n_1)f_2(n_2)\ldots f_g(n_g) \tag{2}$$

and

$$f_i(n_i) = \begin{cases} \dfrac{X_i^{n_i}}{n_i!}, & n_i \leq s_i, \\[2ex] \dfrac{X_i^{n_i}}{s_i!s_i^{n_i-s_i}}, & n_i > s_i; \qquad i=1,\ldots,g. \end{cases} \tag{3}$$

The single-machine, multiple-machine, and infinite-machine groups correspond, respectively, to $s_i = 1$, $1 < s_i < n$, and $s_i \geq n$.

One way to measure system performance is the production rate. Instead, we scale production (via the scaled workload) so that it is independent of actual time, but is a function of scaled time. By using a dimensionless production function, we can get a common base with which to compare allocations. For a system such as that depicted in Figure 1 with n parts in the system, the expected production rate, which is the expected number of parts produced per unit time, can be obtained as a function of g, n, S, and X, that is, as a function of assigned workload, X_i, and grouping, s_i. In fact, for a particular scaling of q_i, the normalized production function, Pr(g,n;S,X), is given by Reiser and Kobayashi (1975) as:

$$Pr(g,n;S,X) = \frac{G(g,n-1;S,X)}{G(g,n;S,X)} . \tag{4}$$

The expected production rate is obtained from the production function by an appropriate scaling factor as follows. The expected departure rate from machine group i is (see Baskett et al. [1975]):

$$\frac{X_i}{t_i}\frac{G(g,n-1;S,X)}{G(g,n;S,X)} \text{ parts per unit time,} \qquad i = 1,\ldots,g.$$

Summing over i, the total production rate of the system (parts per unit time into machine groups) is A Pr(g,n;S,X), where from equation (1),

$$A = \sum_{i=1}^{g} \frac{X_i}{t_i} = m \sum_{i=1}^{g} q_i/(\sum_{j=1}^{g} q_jt_j).$$

Since the parameter A is held constant in what follows, where we optimize over X and S (with m, n, g, q_i, and $\sum_{j=1}^{g} q_jt_j$ fixed), we maximize total expected production rate by maximizing the production function of equation (4), Pr(m,n;S,X), instead.

Any effects that the following assumptions of the model might have on the application of the results to FMSs are now discussed.

1. There are always n parts present.
   During FMS control, there is usually a fixed number of pallets on the system, with a new part loaded immediately after another is unloaded. Having too many parts causes congestion while having too few parts results in an underutilized system.

2. Each machine group i contains a buffer with room for $n-s_i$ waiting parts.
   The "adequate-buffer" assumption eliminates the consideration of the blocking or starving of machines. If each machine has a large enough buffer, this assumption poses no problems. For a very small (or no) buffer, this assumption might seem to be unreasonable. However, robustness of this model has been observed (Suri [1983]).

3. The parts follow a probabilistic routing through the system.
   The expected production rate provides a robust measure despite requiring only the mean number of visits to each machine group (Solberg [1979]).

Further information concerning the relevance of CQN model results to FMS design, planning, and control can be found in Stecke and Solberg (1981a), Cavaillé and Dubois (1982), and Suri (1983). The application of our CQN results to FMS planning problems are discussed in §4.

We now investigate the cases of maximum pooling (g=1), partial pooling (1<g<m), and no pooling (g=m). These achieve lower and lower maximum production rates with the extreme cases:

maximum pooling--$Pr(1,n;(m),(m)) = \min\{1,n/m\}$;

no pooling--$Pr(m,n;(\vec{1}),1^*) = n/(n+m-1)$,

where 1^* is the balanced workload, $X_1^* = X_2^* = \ldots = X_g^* = 1$, that achieves the maximum expected production in the no-pooling case (Stecke and Morin [1982]). From now on, regarding X^* or 1^*, * refers to the optimum (un)balance that maximizes expected production. In addition, $(\vec{1}) = S = (1,1,\ldots,1)$.

2. MAXIMUM POOLING: ONE GROUP

The first case that is examined maximizes flexibility in a manufacturing system: all machines are pooled into one group, so g=1. Each machine can perform any operation. Then, $s_1=X_1=m$. The normalized expected production rate from this system is compared to the best possible normalized expected production rate achieved from a single-machine, g-machine group system.

For any workload X, $f_1(n)$, from equation (3), is given by:

$$f_1(n) = \begin{cases} \dfrac{m^n}{n!}, & n \leq m; \\ \dfrac{m^n}{m!\ m^{n-m}}, & n > m. \end{cases}$$

Consequently, the production function, from equation (4), is

$$Pr(1,n;S,X) = \frac{f_1(n-1)}{f_1(n)}$$

$$= \begin{cases} n/m, & \text{if } n \leq m; \\ 1, & \text{if } n > m. \end{cases}$$

From Stecke and Morin (1982), the optimal solution to the single-machine machine group system (that is, a single-server CQN with g servers) is a balanced workload: $X_1^*=X_2^*=\ldots=X_g^*=1$. The comparison is stated as Theorem 1.

Theorem 1. Except for those systems in which there is only one part, the production rate obtained from a totally pooled system of machines is strictly greater than the best that can be obtained from a system containing no pooling; when n is 1, the expected production rate from the two systems is identical.

Proof: For the system of groups of single machines,

$$\begin{aligned} Pr(m,n;(\vec{1}),X^*) &= n/(n+m-1) \qquad \text{(Stecke and Morin [1982])} \\ &\leq \min\{n/m,1\} \\ &= Pr(1,n;(m),(m)), \end{aligned}$$

which is the production rate for the pooled system. In addition, if there is only one part in the system, then $Pr(m,n;(\vec{1}),X^*) = Pr(1,n;(m),(m)) = 1/m$.

Intuitively, the more pooling the better, with respect to production. Because of technological considerations, however, pooling all machines into one group is not usually possible. For example, in an FMS it is usually impossible to place all of the cutting tools required for all operations to be performed by a particular machine type in only one limited-capacity tool magazine. This implies that some partitioning of the machine tools must occur. The following section analyzes grouping and loading problems for those cases in which only limited pooling is possible.

3. PARTIAL POOLING: SEVERAL GROUPS

This section addresses questions concerning:

i) how to partition m machines into g groups; that is, choose s_i to maximize $Pr(m,n;S,X)$ subject to $\sum_{i=1}^{g} s_i = m$; and

ii) how to allocate a total amount of workload among the system of grouped machines; that is, choose X_i (by choosing average processing times for fixed q_i) to maximize production subject to $\sum_{i=1}^{g} X_i = m$.

The section is organized as follows. We first examine the simplest partially pooled system: three machines in two groups (S = (1,2)). Because of the cumbersome nature of the production function, only small problems with a maximum of three parts can be solved analytically. The analysis is continued numerically and graphically for a system having four parts, on up to a system having fourteen parts (quite saturated for three machines). Finally, the limiting case of a system comprising an infinite number of parts is examined. Next, four-machine systems are investigated in order to demonstrate the pooling, grouping, and loading results. A system consisting of four single machines has been examined in Stecke and Morin (1982). A system consisting of four pooled machines is analyzed here in §2. The remaining cases involving two groups, S = (2,2) and (1,3), and three groups, S = (1,1,2), are presented. Then several seven-machine systems, grouped in different ways, are displayed, compared, and discussed. The general results which we have been successful in proving are given in §3.2. Finally, we discuss the generality of those results whose formal proofs remain elusive, but which have proven valid for the many individual cases that we have examined.

3.1 Analytical, Numerical, and Graphical Results

Several systems are analyzed in order of increasing complexity. The simplest partially pooled system that can be examined is one consisting of two machines in one group and a single machine in the other, as shown in Figure 3.

The output from this system is compared to:
i) a system of three single machines (Stecke and Morin [1982]), and
ii) the system of three pooled machines (§2).

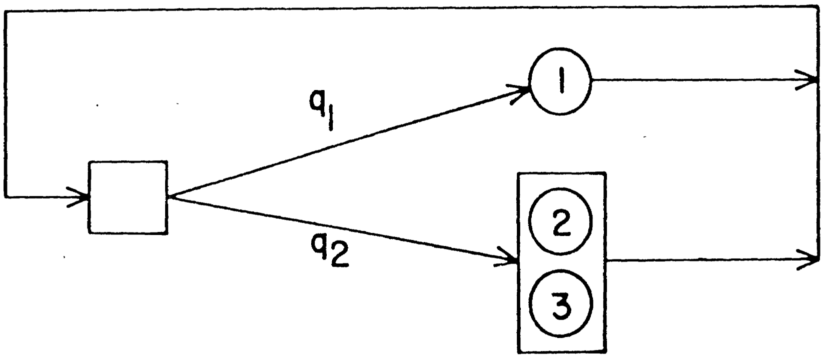

Figure 3. A Three-Machine, Partially Pooled System.

For an $S = (1,2)$ system, $g=2$, $s_1=1$, and $s_2=2$. Then,

$$X_i = \text{normalized workload assigned to machine group } i$$
$$= (s_1+s_2)(q_it_i)/(q_1t_1+q_2t_2)$$
$$= 3q_it_i/(q_1t_1+q_2t_2).$$

From equation (3), $f_1(n_1)$ and $f_2(n_2)$ are:

$$f_1(n_1) = X_1^{n_1}, \qquad \text{for all } n_1;$$

$$f_2(n_2) = \begin{cases} \dfrac{X_2^{n_2}}{n_2!}, & n_2 \leq 2; \\ \dfrac{X_2^{n_2}}{2!2^{n_2-2}=2^{n_2-1}}, & n_2 > 2. \end{cases}$$

$$Pr(2,n;(1,2),X) = \frac{\sum_{n_1+n_2=n-1} f_1(n_1)\cdot f_2(n_2)}{\sum_{n_1+n_2=n} f_1(n_1)\cdot f_2(n_2)}.$$

Analyses of systems with one and two parts can be found in Stecke and Solberg (1981a). For three parts in the system, the normalized production rate is given by:

$$Pr(2,3;(1,2),X) = \frac{X_1^2+X_1X_2+X_2^2/2}{X_1^3+X_1^2X_2+X_1X_2^2/2+X_2^3/4} = \frac{4X_1^2+4X_1X_2^2+2X_2}{4X_1^3+4X_1^2X_2+2X_1X_2^2+X_2^3}.$$

Reducing this expression to a single variable by the elimination of X_1 $(= 3 - X_2)$, we obtain after some algebraic manipulations,

$$Pr(2,3;(1,2),X) = \frac{2(X_2^2-6X_2+18)}{108-72X_2+18X_2^2-X_2^3}.$$

Proposition 2. For the three-machine, two-group system with three parts, the expected production rate is maximized by assigning a unique unbalanced workload per machine to each group.

Proof: $\frac{\partial Pr(2,3;(1,2),X)}{\partial X_2}$

$$= \frac{(108-72X_2+18X_2^2-X_2^3)(4X_2-12)+(2X_2^2-12X_2+36)(72-36X_2+3X_2^2)}{(108-72X_2+18X_2^2-X_2^3)^2}$$

$$= \ldots$$

$$= \frac{2(X_2^4-12X_2^3+90X_2^2-432X_2+648)}{(108-72X_2+18X_2^2-X_2^3)^2} .$$

Setting $\frac{\partial Pr(2,3;(1,2),X)}{\partial X_2} = 0$ implies that $X_2 = 2.38965$, which is the only feasible, real root in the interval [0,3]. Then, 3 - 2.38965 gives $X_1 = .61035$. Pr(2,3;(1,2),(.61, 2.39)) = .7472. Therefore the maximum normalized expected production rate, .7472, occurs at an unbalanced workload per machine, that is, at $X_1 = .610$ and $X_2/2 = 1.195$.

This maximum production rate is better than Pr(2,3;(1,2),(1,2)) = .7143, which is the balanced assignment's scaled production rate. This illustration proves that production is increased significantly when the workload per machine for a particular system of grouped machines of unequal sizes is deliberately unbalanced. That is, the performance of the unbalanced system is better than that of a balanced system.

Mathematical analysis is useful only for small problems. The expansion of Pr(g,n;S,X) becomes too cumbersome to perform manually because of the exponentially increasing number of states. However, the function can be evaluated quickly by using Buzen's algorithm (1973). A computer program called CAN-Q (Solberg [1980]), that evaluates multi-server CQNs, was adapted (Stecke and Solberg [1981a]) to accept our particular problem parameters. For a system of grouped machines, the program efficiently calculates:

i) the maximum production;
ii) the balanced production;

iii) the percentage increase in the maximum over the balanced production;
iv) the optimal loading with respect to maximum production;
v) the percentage decrease in the workload per machine assigned to the first machine group.

The computer program was used to evaluate Pr(2,n;(1,2),X) for 400 consecutive points (X_1,X_2) for X_1 in [0,3] and X_2 in [1.5,0]. Figure 4 is a plot of the 400 points $(X_1,$Pr(2,n;(1,2),X)) for twelve production functions: n=4,5,...,14, and the limiting case of n equal to infinity. The vertical line through $X_1=1$ cuts each production curve at the point of a balanced assignment, that is, where $X_1=X_2/2=1$. The crosses mark the points of maximum value. The following information is extracted from the curves of Figure 4:

i) The functions, Pr(2,n;(1,2),X), are strongly quasiconcave for n=4,5,...,14 and infinity. This indicates the uniqueness of the global maximum. (See Stecke [1983b].)

ii) The production functions are not symmetric. If they were, then from Stecke and Morin (1982), the balanced point would be optimal.

iii) The unique, maximum production rate is achieved by underutilizing (overutilizing) the smaller (larger) group of machines. For each finite n, the maximum is strictly larger than the balanced.

iv) As the number of parts in the system increases, the degree of unbalance of the optimal allocation decreases. For a totally saturated system (n=∞), the maximum expected production rate is the same as for the balanced system.

Table I provides values of the balanced (Pr(Bal)) and maximum (Pr(Max)) production rates, and the optimal allocation per machine in each group. Note that with only three machines, six or seven parts begin to saturate the system.

For a totally pooled system, Pr(1,n;(3),X) = 1 for all $n \geq 2$, which indicates that fewer groups are better. Similar evidence is found in Table I by comparing the maximum (which is also the balanced) production rate achieved by three single machines to the partially pooled system. In particular, for each finite $n > 2$,

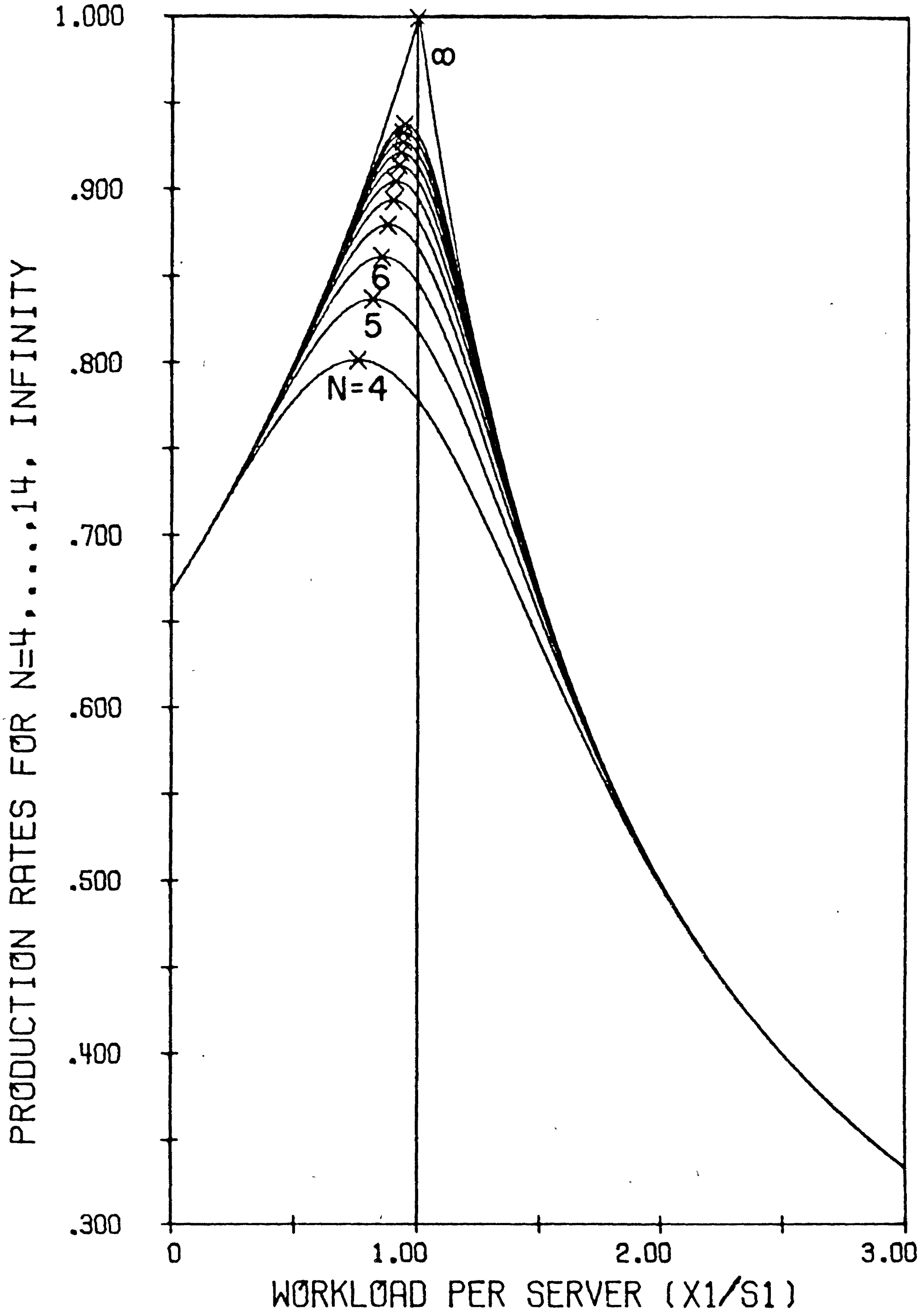

Figure 4. Expected Production Rate as a Function of Workload for Three Machines: S = (1,2).

TABLE I

Three Machines:
Balanced Versus Maximum Production Rates and Optimal Workloads

a Two Groups of One and Two Machines: S = (1,2)

n	Pr(Bal)$^+$	Pr(Max)**	Percent Increase in Prod Rate	Percent Decrease in Wkld on Mac 1	X_1^*/s_1	X_2^*/s_2
4	.777	.801	3.15	24	.757	1.121
5	.817	.837	2.38	18	.817	1.091
6	.845	.861	1.92	15	.855	1.072
7	.856	.880	1.61	12	.877	1.061
8	.881	.894	1.39	10	.900	1.050
9	.894	.905	1.23	9	.907	1.046
10	.904	.914	1.11	8	.922	1.039
11	.912	.921	1.01	7	.930	1.035
12	.919	.927	.93	6	.937	1.031
13	.925	.933	.86	6	.937	1.031
14	.930	.937	.80	5	.945	1.027
∞	1.000	1.000	0	0	1.000	1.000

b Three Groups of One Machine Each: S = (1,1,1)

n	Pr(Max) and Pr(Bal)
4	.667
5	.714
6	.750
7	.777
8	.800
9	.818
10	.833
11	.846
12	.857
13	.867
14	.875
∞	1.000

$^+$Pr(Bal) = $Pr(g,n;S,X/S = \vec{1})$

**Pr(Max) = $Pr(g,n;S,X^*)$

$$\Pr(\text{Max})(3,n;(1,1,1),1) < \Pr(2,n;(1,2),\text{Max or Bal } X) < \Pr(1,n;(3),X).$$

For four-machine systems, all possible groupings include:

i) one group: S = (4);
ii) two groups: S = (1,3) and S = (2,2); see Figures 5 and 6, respectively;
iii) three groups: S = (1,1,2); see Stecke (1981);
iv) four groups: S = (1,1,1,1); see Stecke and Morin (1982).

First, two systems of two groups are compared. All of the production functions are explicitly quasiconcave. For the equal-sized groups ((2,2)--Figure 6), the functions are symmetric and the balanced point is optimal. For unequal-sized groups ((1,3)--Figure 5), the functions are not symmetric and the balanced point is not optimal. Also, note from Table II that the grouping S = (1,3) is better than S = (2,2) and, in addition, that Pr(Max)(1,3) is significantly higher than Pr(Bal/Max)(2,2), which is only slightly better than Pr(Bal)(1,3) for all finite n. These observations demonstrate the grouping and loading results that were summarized in the introduction. A reason for these unbalancing phenomena is that a larger group of pooled machines is more efficient (i.e., can complete more jobs per unit time), and can hence be assigned more than its "fair share" (i.e., balanced) amount of work.

To illustrate the results, notice from Table II that when there are four parts in the system, say, the lone machine should be assigned a proportion of only .5 units of work on the average, while each of the other three machines should be assigned 1.167 units. This unbalanced workload per machine increases production rate by 8.23 percent. Tables II and III show that both configurations S = (1,3) and (2,2) produce larger maximum production rates than system S = (1,1,2). Finally, Table IV shows the optimality of a balanced workload for a system of four single machines. Note that for all finite n, Pr(Max)(1,1,1,1) < Pr(Max)(1,1,2).

The analysis is similar for systems of five, six, seven, and fifty machines (see Stecke and Solberg [1981a]). For example, for seven machines in two groups (summarized in Figures 7 and 8 and Table V):

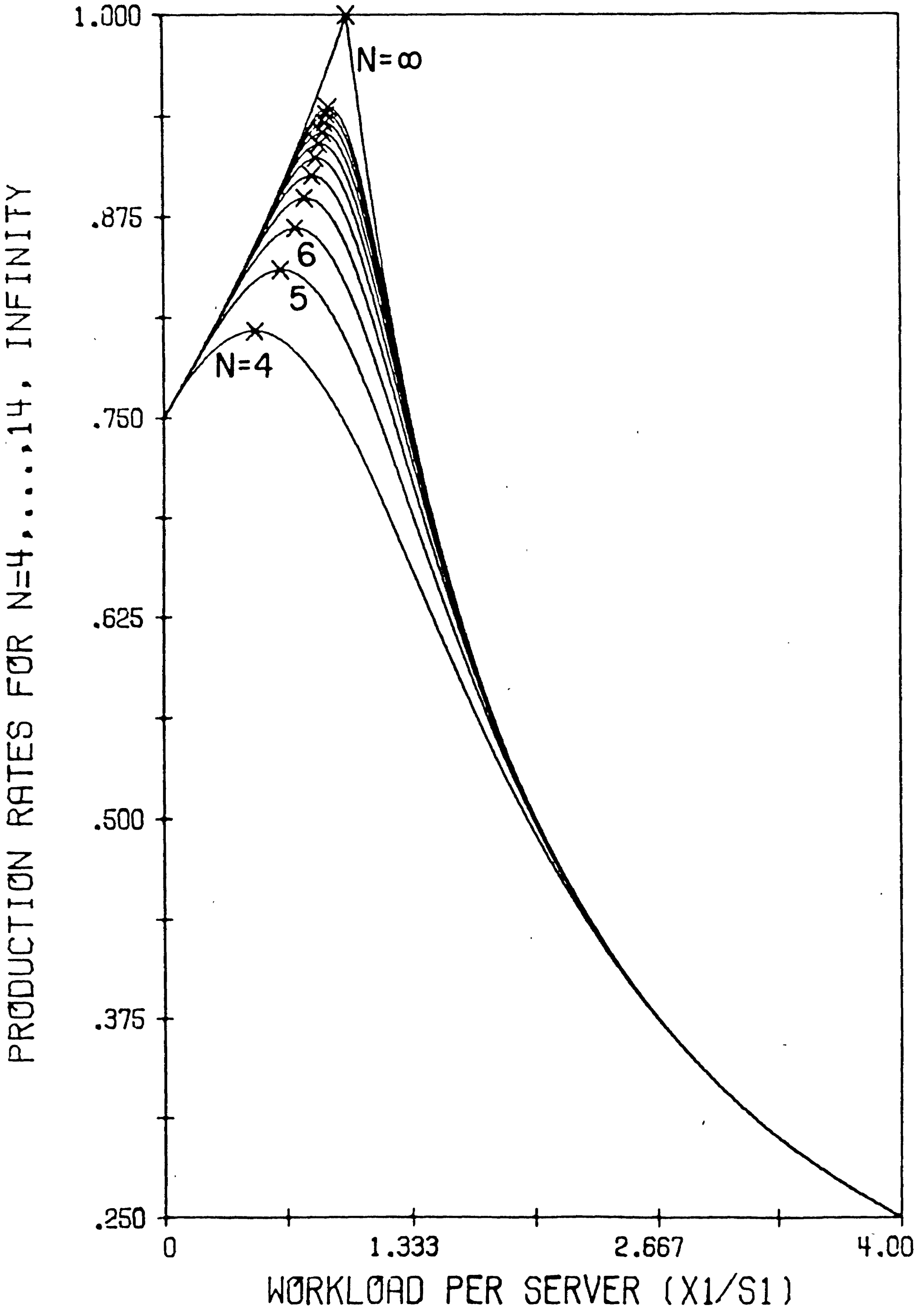

Figure 5. Two Groups of Sizes One and Three.

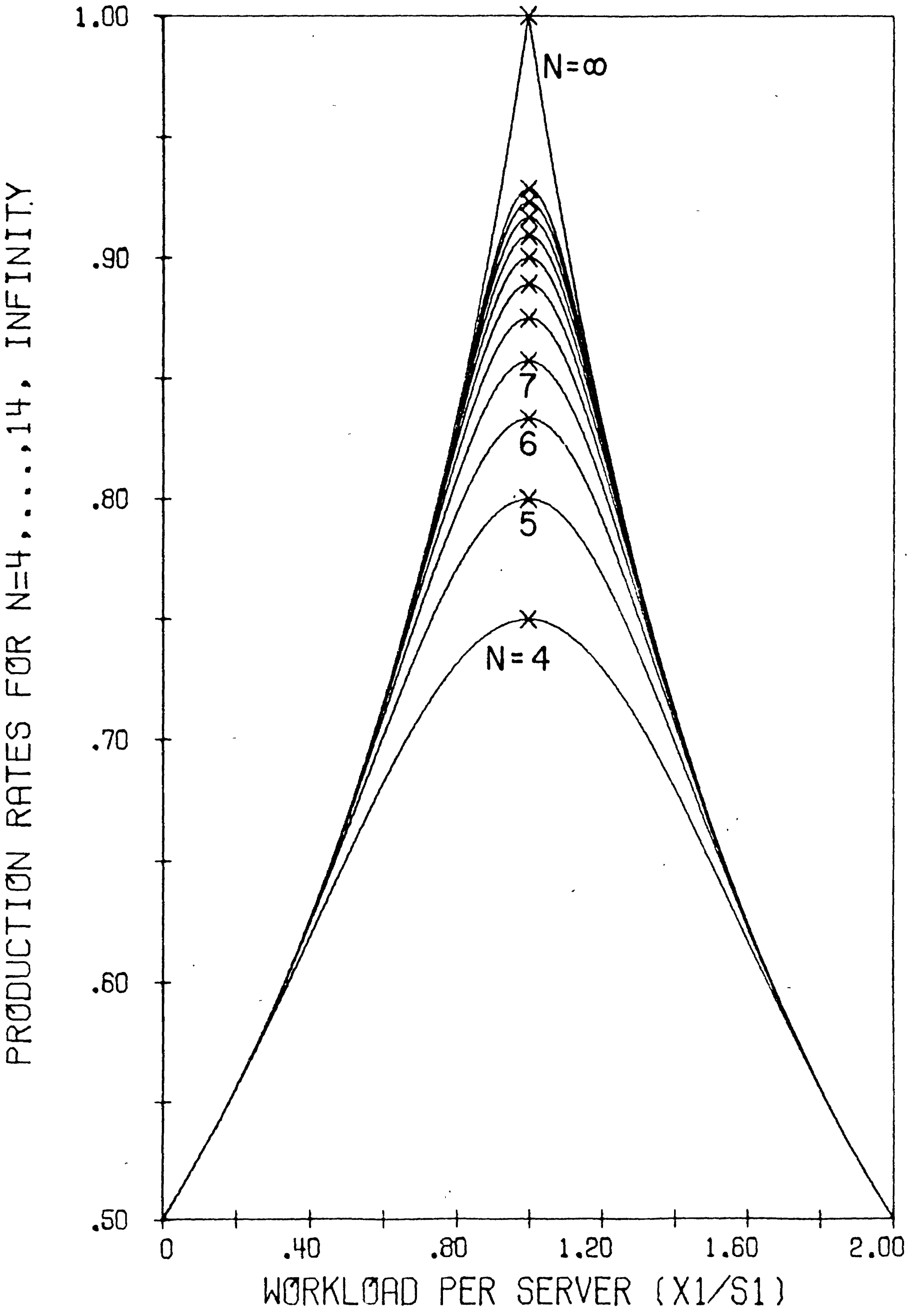

Figure 6. Two Groups of Sizes Two and Two.

TABLE II

Four Machines:
Balanced Versus Maximum Production Rates and Optimal Workloads

a Two Groups of One and Three Machines: S = (1,3)

n	Pr(Bal)	Pr(Max)	Percent Increase in Prod Rate	Percent Decrease on Mac 1	X_1^*/s_1	X_2^*/s_2
4	.743	.804	8.23	50	.500	1.167
5	.795	.842	5.87	36	.640	1.120
6	.830	.868	4.56	28	.720	1.093
7	.855	.887	3.72	23	.770	1.077
8	.873	.901	3.14	19	.810	1.063
9	.887	.912	2.71	17	.830	1.057
10	.889	.920	2.38	15	.850	1.050
11	.908	.927	2.13	13	.870	1.043
12	.916	.933	1.92	12	.880	1.040
13	.922	.939	1.75	11	.890	1.037
14	.928	.943	1.61	10	.900	1.033
∞	1.000	1.000	0	0	1.000	1.000

b Two Groups of Two Machines Each: S = (2,2)

n	Pr(Bal)	Pr(Max)	Percent Increase in Prod Rate	Percent Decrease on Mac 1	X_1^*/s_1	X_2^*/s_2
4	.750	.750	0	0	1.00	1.00
5	.800	.800	0	0	1.00	1.00
6	.833	.833	0	0	1.00	1.00
7	.857	.857	0	0	1.00	1.00
8	.875	.875	0	0	1.00	1.00
9	.889	.889	0	0	1.00	1.00
10	.900	.900	0	0	1.00	1.00
11	.909	.909	0	0	1.00	1.00
12	.917	.917	0	0	1.00	1.00
13	.923	.923	0	0	1.00	1.00
14	.929	.929	0	0	1.00	1.00
∞	1.000	1.000	0	0	1.00	1.00

TABLE III

Four Machines in Three Groups: S = (1,1,2)

n	Pr(Bal)	Pr(Max)	Percent Increase in Prod Rate	Percent Decrease on Mac 1	X_1^*/s_1	X_2^*/s_2
4	.640	.665	3.81	24.00	.760	1.240
5	.695	.715	2.89	17.33	.827	1.173
6	.735	.752	2.32	14.00	.860	1.140
7	.766	.781	1.93	11.33	.887	1.113
8	.790	.804	1.65	10.00	.900	1.100
9	.810	.822	1.44	8.67	.913	1.087
10	.827	.837	1.27	7.33	.927	1.073
11	.840	.850	1.14	6.67	.933	1.067
12	.852	.861	1.03	6.00	.940	1.060
13	.863	.871	.94	5.33	.947	1.053
14	.872	.879	.86	5.33	.947	1.053
∞	1.000	1.000	0	.00	1.000	1.000

TABLE IV

Four Single Machines: S = (1, 1, 1, 1)

n	Balanced Prod Rate	Maximum Prod Rate	X_1*	X_2*
4	.571	.571	1.00	1.00
5	.625	.625	1.00	1.00
6	.667	.667	1.00	1.00
7	.700	.700	1.00	1.00
8	.727	.727	1.00	1.00
9	.750	.750	1.00	1.00
10	.769	.769	1.00	1.00
11	.786	.786	1.00	1.00
12	.800	.800	1.00	1.00
13	.812	.812	1.00	1.00
14	.824	.824	1.00	1.00
∞	1.000	1.000	1.00	1.00

$$\Pr(\mathrm{Max})(1,6) \gg \Pr(\mathrm{Max})(2,5) > \Pr(\mathrm{Max})(3,4) > \Pr(\mathrm{Bal})(3,4)$$
$$> \Pr(\mathrm{Bal})(2,5) > \Pr(\mathrm{Bal})(1,6).$$

3.2 General Results

First, the optimality of balanced workloads when all machine group sizes are equal is proven. Definitions of generalized concavity and symmetric functions can be found in Berge (1963) and Stecke and Solberg (1981a). Proofs can be found in Stecke and Solberg (1981a) and are direct generalizations of those in Stecke and Morin (1982); hence they are not repeated here.

Theorem 3. If each group contains the same number of machines ($s_i = s$, for all i), the expected production rate is maximized when the allocation per machine is balanced, that is, $X_i/s = 1$, $i = 1,\ldots,g$.

Proof: The set of feasible loadings, χ, is closed and S-convex (Stecke and Solberg [1981a], Stecke and Morin [1982]).

Pr(g,n;S,X) is a symmetric function (Stecke and Solberg [1981a], Stecke and Morin [1982]).

Pr(g,n;S,X) is strictly quasiconcave, but not concave (Stecke [1983b]).

By symmetry and quasiconcavity, Pr(g,n;S,X) is S-concave (Berge [1963]).

The set, χ^*, of points maximizing Pr(g,n;S,X) over the set χ is a closed, S-convex set (Greenberg and Pierskalla [1970]).

χ^* is not empty since Pr(g,n;S,X) ε [0,1] (i.e., is bounded) for all g and n, and for $\chi \; \varepsilon \; [0,gs]$.

The symmetric point of χ is the point [s,s,...,s] (by definition).

The symmetric point [s,...,s] $\varepsilon \; \chi^*$ (Greenberg and Pierskalla [1970]).

Therefore, a balanced allocation per machine maximizes the expected production rate, or $X_i/s = 1$, $\quad i = 1,\ldots,g$.

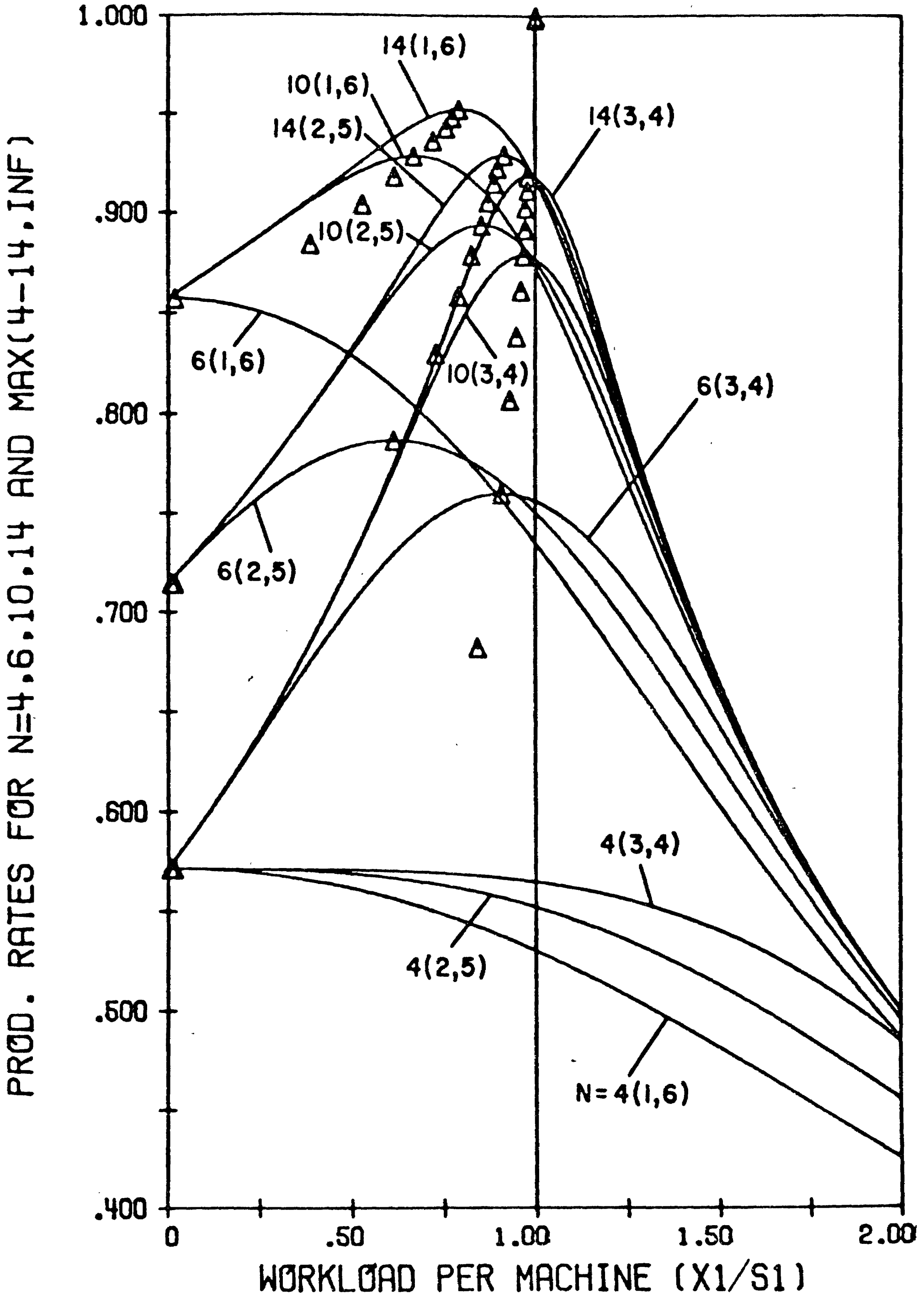

Figure 7. Expected Production Rates for Seven Machines in Two Groups: (1,6), (2,5), and (3,4).

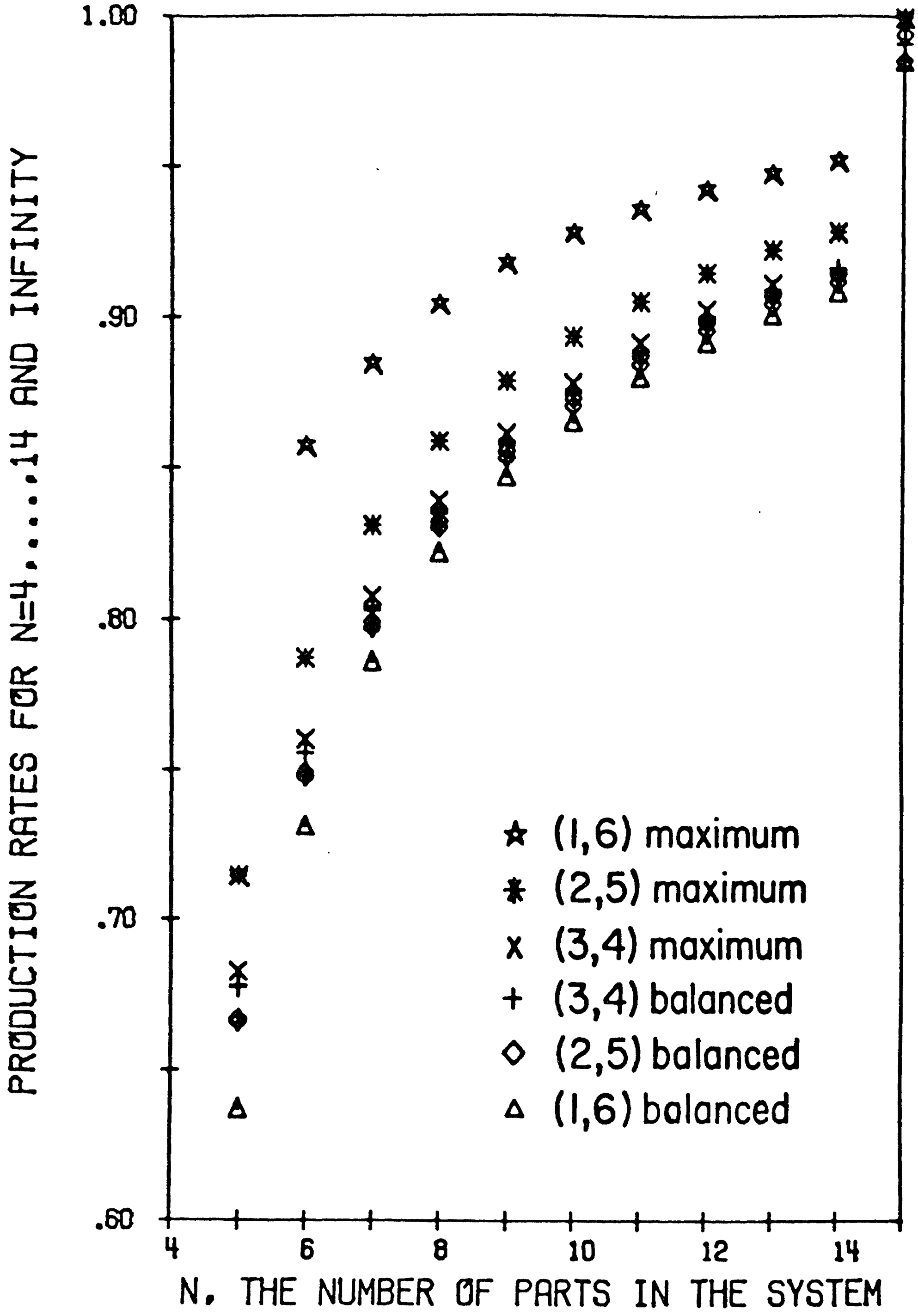

Figure 8. Maximum Versus Balanced: (1,6), (2,5), (3,4).

TABLE V

Seven Machines:
Balanced Versus Maximum Production Rates and Optimal Workloads

a Two Groups of One and Six Machines: S = (1,6)

n	Pr(Bal)	Pr(Max)	Percent Increase in Prod Rate	Percent Decrease in Wkld on Mac 1	X_1^*/s_1	X_2^*/s_2
4	.529	.571	7.98	100	.000	1.164
5	.637	.714	12.07	100	.000	1.164
6	.731	.857	17.19	100	.000	1.164
7	.786	.885	12.53	62	.385	1.102
8	.822	.905	10.05	48	.525	1.079
9	.847	.918	8.40	39	.612	1.064
10	.866	.929	7.24	34	.665	1.055
11	.880	.937	6.37	28	.717	1.047
12	.892	.943	5.70	25	.752	1.041
13	.901	.948	5.18	23	.770	1.038
14	.909	.952	4.75	13	.787	1.035
∞	1.000	1.000	.00	00	1.000	1.000

b Two Groups of Two and Five Machines: S = (2,5)

4	.552	.571	3.52	100	.000	1.396
5	.667	.714	7.09	100	.000	1.396
6	.749	.787	5.04	39	.612	1.155
7	.799	.830	3.95	27	.726	1.109
8	.832	.859	3.23	21	.787	1.085
9	.856	.879	2.72	18	.822	1.071
10	.874	.894	2.36	15	.849	1.060
11	.887	.906	2.08	13	.866	1.053
12	.898	.915	1.87	12	.884	1.046
13	.907	.923	1.70	11	.892	1.043
14	.915	.929	1.56	9	.910	1.036
∞	1.000	1.000	.00	0	1.000	1.000

c Two Groups of Three and Four Machines: S = (3,4)

4	.565	.571	1.14	100	.000	1.756
5	.678	.682	.69	16	.840	1.120
6	.756	.760	.54	10	.904	1.072
7	.804	.807	.44	7	.927	1.054
8	.836	.838	.36	6	.945	1.041
9	.859	.861	.31	4	.957	1.032
10	.876	.878	.27	4	.962	1.028
11	.890	.892	.25	3	.968	1.024
12	.901	.902	.22	3	.968	1.024
13	.910	.911	.21	3	.974	1.019
14	.917	.918	.20	3	.974	1.019
∞	1.000	1.000	.00	0	1.000	1.000

The nonsymmetry of the production function for systems with groups of different sizes has been observed in §3.1.

Theorem 4. For a system in which there are at least two groups of different sizes and $n > \min_i s_i$, $Pr(g,n;S,X)$ is not symmetric in X.

The proof is straightforward, but tedious, and can be found in Stecke and Solberg (1982).

Conjecture 5. Given a system of groups of pooled machines of unequal sizes, the production function is maximized by a unique unbalanced allocation of the workload per machine. In particular, more (less) than the balanced amount of work per machine is assigned to the larger (smaller) groups of machines.

The validity of Conjecture 5 has been observed for all of the systems we have examined. The generality of the assertion is conjectured but not proven to date because of the numerical nature of the production function.

There is an interesting relation between a balanced and maximum allocation in the limiting case:

Conjecture 6. In a totally saturated system ($n=\infty$), the optimal allocation is balanced, even for unequally-sized machine groups.

Corollary 7. For a totally saturated system ($n=\infty$), the maximum expected production rate is the same as the production rate for the balanced system, that is,

$$\lim_{n\to\infty} \underset{X}{\text{maximum}}\ Pr(g,n;S,X) = 1 = \lim_{n\to\infty} Pr(g,n;S,\vec{1}) \text{ for any g and S.}$$

The following conjecture and associated corollary define how to best group m machines into g groups. Throughout, groups are ordered according to increasing size, that is, $s_1 \leq s_2 \leq \ldots \leq s_g$. When machines are shifted into different groups, the ordering according to sizes is preserved. A general statement is first provided which all of our investigations indicate to be valid.

Conjecture 8. Suppose there are m machines, g groups, and s_i machines in group i, and groups are ordered so that $s_1 \leq s_2 \leq \dots \leq s_g$. Then, for any integer $K > 0$, we have that:

i) $\underset{X}{\text{maximum}}\ Pr(g,n;(s_1,\dots,s_i+K,\dots,s_g-K),X)$ is strictly less than $\underset{X}{\text{maximum}}\ Pr(g,n;(s_1,\dots,s_i,\dots,s_g),X)$;

ii) $Pr(g,n;(s_1,\dots,s_i+K,\dots,s_g-K),S)$ is strictly greater than $Pr(g,n;(s_1,\dots,s_g),S)$;

iii) $\max_X Pr(g,n;(s_1,\dots,s_i,\dots,s_j,\dots,s_g),X)$ is strictly greater than $\max_X Pr(g,n;(s_1,\dots,s_i+K,\dots,s_j',\dots,s_g),X)$, for $s_i+K = s_{i+1}+K = \dots = s_j'$.

The order of groups according to size is preserved throughout. In particular, $s_i + K \leq \dots \leq s_g - K$.

The first and third parts of Conjecture 8 state that the maximum expected production rate is larger for more unbalanced system configurations of pooled machines. Similarly, the second part indicates that equalizing group sizes decreases the maximum expected production rate. Finally, if the allocation per machine is balanced, then group sizes should be equal to maximize expected production.

The following special case demonstrates Conjecture 8. If m machines are to be partitioned into two groups, with x machines in group 1 and y machines in group 2, all partitions can be sequenced according to the maximum expected production rate.

Corollary 9. All partitions of m machines into two groups are ordered according to the maximum expected production rate as follows:

$$\begin{array}{c}(1,\ m\text{-}1)\\ (2,\ m\text{-}2)\\ (3,\ m\text{-}3)\\ \vdots \\ \left(\begin{array}{ll}(m\text{-}1)/2,\ (m\text{+}1)/2, & \text{if } m \text{ is odd;}\\ m/2,\ m/2, & \text{if } m \text{ is even.}\end{array}\right)\end{array}$$

Additional interesting properties of the production function can be found in Stecke and Solberg (1981a) and Stecke and Schmeiser (1983).

4. DISCUSSION

The unbalanced grouping result is not initially intuitively obvious. To see this, suppose that a new machine tool is purchased and could be pooled with either group of an existing three-machine system, S = (1,2). Since there are advantages (and diminishing returns) to pooling, one might think it best to include the new machine tool with the single-machine group, to obtain system (2,2). If the workload per machine were to be balanced, then the expected production rate from system (2,2) would be slightly higher than that of the alternative two-group system (1,3) (see Table II). However, since the expected production rate of a (1,3) system is maximized with an unbalanced workload per machine and also is higher than that of a (2,2) system, then one would attempt to include the new machine with the second, larger group.

The optimality of unbalanced allocations for unbalanced partitions of pooled machines can also initially seem to be counterintuitive. The most heavily utilized machine is traditionally referred to as a bottleneck machine; the idea is that operations which require the bottleneck machine tend to wait longer in the queue to be processed. The practice in industry is to shift work to lessen the load on the bottleneck machine, which is called load-leveling.

However, our results indicate that a machine could be critical to overall system behavior even if it were assigned less work than another machine on the average. A machine could be assigned less work than another, and still be a "bottleneck," not in the usual sense of the word, but in that system peformance could improve by assigning even less work to this lower assigned machine. In addition, the advantages of pooling are over and above a savings realized in the travel time required to get from machine to machine.

The application of these results to flexible manufacturing can now be described. Suppose that there is a set of operations, whose total workload has been scaled to equal m_j, that require a particular machine type j. (There are m_j machines of type j). The best possible solutions to the FMS grouping and loading problems would be that all m_j machines are pooled into one group and hence identically tooled. Then each machine would be capable of performing all operations. However, this maximum pooling situation is rarely possible since the cutting tools required for all of the operations cannot usually fit into each machine tool's limited-capacity tool magazine. If g_j tool magazines are required to hold all required tools, then the m_j machines have to be partitioned into at least g_j groups. The solution to the continuous FMS loading problem is a set of optimal allocation ratios, X^*_{ij}, $i = 1,\ldots,g_j$, or ratios at which the machine groups of type j should be loaded to maximize expected production. Operations should then be assigned to the machine groups so that total operation time/machine would be as near as possible to the ratios: X^*_{ij}/s_{ij}, rather than balanced.

At present, there are no known simple conditions (of an operationally useful nature) by which to characterize the optimum unbalanced solutions. Speculations that the production function would be optimized at a point where the probabilities of finding an available server are equal, or where the mean queue lengths per server are equal, or several other possibilities, are not supported by our computational experiments. The optimal allocation ratios to maximize expected production are found by searching over the possible workload parameters.

A larger group of pooled machines can be loaded more heavily simply because pooled servers are more efficient than single servers. In addition, a larger group automatically increases flexibility by increasing the number of job routes. Production of more than one part type, each with alternate routes, increases system variability. Pooling machines helps a system automatically adapt to congestion (in part caused by increased by system variability) and machine/cutting tool/cart breakdowns.

The benefits of pooling machines can be obtained only in conjunction with an on-line control strategy, where decisions are based on the present state of the system. A CQN model was chosen in order to approximate real-time control capabilities via the probablistic routing and also to consider queueing and congestion in an FMS. Because of the flexibility available in automated manufacturing and the use of computers, real-time scheduling can become a viable and welcome alternative to a fixed, static, off-line-generated schedule.

This paper reports a study of balancing versus unbalancing, with particular applications to FMS grouping and loading problems. However, additional loading objectives have been suggested that also proved superior to balancing when applied using a detailed simulation of, and data from, an existing FMS (Stecke and Solberg [1981b]). The superior performance was surprising because:

i) the resultant system was extremely unbalanced; and

ii) there was no pooling involved.

In addition, pooling objectives performed well.

5. DIRECTIONS FOR FUTURE RESEARCH

One goal is to apply these theoretical results to real FMS design, planning, and control problems. To increase applicability, the existing analytic models can be further refined and extended. For example, CQN methodology allows multiple part types, each with its own branching probabilities, provided that each FCFS server has the same exponential service time distribution for all part types that visit it (Baskett et al. [1975]). Since it is impractical to compute the normalization constant, G(g,n;S,X), for more than four or five part types, it may be preferable to use an approximate mean value analysis (MVA) to evaluate such a situation (Bard [1979], Chandy and Neuse [1982]). In fact, some FMS modeling using MVA to develop heuristic algorithms has been done (Hildebrant [1980] and Cavaillé and Dubois [1982]). However, at present, MVA is applicable only to single-server, load-independent systems, with possible heuristic extensions to the multi-server models examined here.

In addition, more detailed models, such as mathematical and simulation models, can be developed. Some model development has been accomplished. Both the grouping and loading problems, with several loading objectives, have been formulated in detail as nonlinear mixed integer problems (Stecke [1983a]). The optimal allocation ratios X_i^*, characterized in this paper, serve as objective function coefficients of detailed loading problems having unbalancing objective functions. Several methods to linearize the nonlinearities were applied using data from an existing FMS. However, the resultant linearized problems can be quite large and time-consuming to solve. For this reason, an efficient optimum-producing algorithm was developed (Berrada and Stecke [1983]). Further research is required to reduce the problems to a consistently manageable size for FMS application. For example, fast, efficient, and good heuristics should be developed to solve grouping and loading problems.

Finally, in an effort to apply the theoretical results presented in this paper to the detailed formulations of Stecke (1983a), a hierarchical approach to solving realistic grouping and loading problems is provided in Stecke (1982).

In addition to planning problems, several control issues have been investigated for FMSs. Kimemia and Gershwin (1978) provide models for determining the operation sequence prior to the release of jobs to the system. Although such methods are applicable to the control of automated transfer lines, they limit flexibility in an FMS. Alternatively, Buzacott (1982) suggests delaying the sequencing decision until either:

i) the job is to be released to the system; or

ii) the moment such a decision is required, i.e., when a machine becomes free. (Real-time control of an FMS helps maximize flexibility.)

Buzacott also provides various operating rules.

The analytical and mathematical models can be used to generate alternative loadings and groupings. However, the "best" objective is highly system-dependent, and the various available alternatives should be evaluated. Every FMS should employ a detailed self-simulation to use with mathematical models to help determine appropriate planning and control strategies prior to their implementation.

ACKNOWLEDGMENT

The research reported here has been partially supported by the National Science Foundation under GRANT No. APR75 15256. The authors are grateful to an anonymous referee for an unusually thorough and thought-provoking review.

REFERENCES

BARD, YONATHAN, "Some Extensions to Multiclass Queueing Network Analysis", Performance of Computer Systems, M. Arato, A. Butrimenko, and E. Gelenbe (eds.), North Holland, Amsterdam (1979).

BASKETT, FOREST, CHANDY, K. MANI, MUNTZ, RICHARD R. and PALACIOS, FERNANDO G., "Open, Closed, and Mixed Networks of Queues with Different Classes of Customers", Journal of the Association for Computing Machinery, Vol. 22, No. 2, pp. 248-260 (April 1975).

BERGE, CLAUDE, Topological Spaces, The Macmillan Company, New York NY (1963).

BERRADA, MOHAMMED and STECKE, KATHRYN E., "A Branch and Bound Approach for Machine Loading in Flexible Manufacturing Systems", Working Paper No. 329, Division of Research, The University of Michigan, Ann Arbor MI (April 1983).

BUZACOTT, JOHN A., "'Optimal' Operating Rules for Automated Manufacturing Systems", IEEE Transactions on Automatic Control, Vol. AC-27, pp. 80-86 (February 1982).

BUZACOTT, JOHN A. and SHANTHIKUMAR, J. GEORGE, "Models For Understanding Flexible Manufacturing Systems", AIIE Transactions, Vol. 12, No. 4, pp. 339-350 (December 1980).

BUZEN, JEFFREY P., "Queueing Network Models of Multiprogramming", Ph.D. Thesis, Harvard University, Cambridge MA (1971).

BUZEN, JEFFREY P., "Computational Algorithms for Closed Queueing Networks with Exponential Servers", Communications of the Association for Computing Machinery, Vol. 16, No. 9, pp. 527-531 (September 1973).

CAVAILLÉ, JEAN-BERNARD and DUBOIS, DIDIER, "Heuristic Methods Based on Mean-Value Analysis for Flexible Manufacturing Systems Performance Evaluation", Proceedings, IEEE Conference on Decision and Control, Orlando FL, pp. 1061-1065 (December 1982).

CHANDY, K. MANI and NEUSE, DOUG, "Linearizer: A Heuristic Algorithm for Queueing Network Models of Computer Systems", Communications of the Association for Computing Machinery, Vol. 25, No. 2, pp. 126-134 (February 1982).

EL-RAYAH, T. E., "The Efficiency of Balanced and Unbalanced Production Lines", International Journal of Production Research, Vol. 17, No. 1, pp. 61-75 (January 1979).

GIAMMO, T., "Validation of a Computer Performance Model of the Exponential Queueing Network Family", Acta Informatica, Vol. 17, No. 2, pp. 137-152 (1976).

GREENBERG, HARVEY J. and PIERSKALLA, WILLIAM P., "Symmetric Mathematical Programs", Management Science, Vol. 16, No. 5, pp. 309-312 (January 1970).

HILDEBRANT, RICHARD R., "Scheduling and Control of Flexible Machining Systems When Machines Are Prone to Failure", Ph.D. Thesis, M.I.T., Cambridge MA (August 1980).

HILLIER, FREDERICK S. and BOLING, RONALD W., "The Effect of Some Design Factors on the Efficiency of Production Lines with Variable Operation Times", Journal of Industrial Engineering, Vol. 17, No. 12, pp. 657-658 (December 1966).

HILLIER, FREDERICK S. and BOLING, RONALD W., "Finite Queues in Series with Exponential or Erlang Service Times: A Numerical Approach", Operations Research, Vol. 15, No. 2, pp. 286-303 (March-April 1967).

HILLIER, FREDERICK S. and BOLING, RONALD W., "On the Optimal Allocation of Work in Symmetrically Unbalanced Production Line Systems with Variable Operation Times", Management Science, Vol. 25, No. 8, pp. 721-728 (August 1979).

HUGHES, P. H. and MOE, G., "A Structural Approach to Computer Performance Analysis", in Proceedings, AFIPS National Computer Conference, Vol. 42, AFIPS Press, Montvale NJ, pp. 109-119 (1973).

IGNALL, EDWARD J., "A Review of Assembly Line Balancing", Journal of Industrial Engineering, Vol. 16, No. 4, pp. 43-52 (July-August 1965).

KELLY, FRANK P., Reversibility and Stochastic Networks, John Wiley & Sons, Inc., New York NY (1979).

KIMEMIA, JOSEPH and GERSHWIN, STANLEY B., "Multicommodity Network Flow Optimization in Flexible Manufacturing Systems", Report No. ESL-FR-834-2, M.I.T., Cambridge MA (September 1978).

KLEINROCK, LEONARD, Queueing Systems, Volume 2: Computer Applications, John Wiley & Sons, New York NY (1976).

KOBAYASHI, HISASHI and REISER, MARTIN, "On Generalization of Job Routing Behavior in a Queueing Network Model", IBM Research Report RC 5252, IBM Thomas J. Watson Research Center, Yorktown Heights NY (February 1975).

LIPSKY, L. and CHURCH, J. D., "Applications of a Queueing Network Model for a Computer System", Computing Surveys, Vol. 9, pp. 205-221 (1977).

MAGAZINE, MICHAEL J. and SILVER, G. L., "Heuristics for Determining Output and Work Allocations in Series Flow Lines", International Journal of Production Research, Vol. 16, No. 3, pp. 169-182 (May 1978).

MAKINO, T., "On the Mean Passage Time Concerning Some Queueing Problems of the Tandem Type", Journal of the Operations Research Society of Japan, Vol. 7, pp. 17-47 (1964).

RAO, NORI PRAKASA, "A Generalization of the 'Bowl Phenomenon' in Series Production System", International Journal of Production Research, Vol. 14, pp. 437-443 (1976).

REISER, MARTIN and KOBAYASHI, HISASHI, "Queueing Networks with Multiple Closed Chains: Theory and Computational Algorithms", IBM Journal of Research and Development, Vol. 19, pp. 283-294 (1975).

REISER, MARTIN and KOBAYASHI, HISASHI, "On the Convolution Algorithm for Separable Queueing Networks", in Proceedings, International Symposium on Computer Performance Modeling, Measurement, and Evaluation, Harvard University, Cambridge MA, pp. 283-294 (1976).

ROSE, C. A., "Validation of a Queueing Model with Classes of Customers", in Proceedings, International Symposium on Computer Performance Modeling, Measurement, and Evaluation, Harvard University, Cambridge MA, pp. 318-325 (1976).

ROSE, C. A., "A Measurement Procedure for Queueing Network Models of Computer Systems", Computing Surveys, Vol. 10, pp. 263-280 (1978).

SECCO-SUARDO, G., "Optimization of a Closed Network of Queues", Report No. ESL-FR-834-3, Electronic Systems Laboratory, M.I.T., Cambridge MA (September 1978).

SHANTHIKUMAR, J. GEORGE and BUZACOTT, JOHN A., "Open Queueing Network Models of Dynamic Job Shops", Working Paper No. 79-024, Department of Industrial Engineering, University of Toronto, Ontario, Canada (August 1979).

SOLBERG, JAMES J., "A Mathematical Model of Computerized Manufacturing Systems", in Proceedings, 4th International Conference on Production Research, Tokyo, Japan (August 1977).

SOLBERG, JAMES J., "Stochastic Modeling of Large Scale Transportation Networks", Report No. DOT-ATC-79-2, Purdue University, W. Lafayette IN (June 1979).

SOLBERG, JAMES J., "CAN-Q User's Guide", Report No. 9 (Revised) NSF GRANT No. APR74 15256, School of Industrial Engineering, Purdue University, W. Lafayette IN (July 1980).

SPIRN, JEFFREY, R., "Queueing Networks with Random Selection for Service", IEEE Transactions on Software Engineering, Vol. 5, No. 3, pp. 287-289 (May 1979).

STECKE, KATHRYN E., "Experimental Investigation of a Computerized Manufacturing System", Master's Thesis, Purdue University, W. Lafayette IN (December 1977).

STECKE, KATHRYN E., "Production Planning Problems for Flexible Manufacturing Systems", Ph.D. Dissertation, Purdue University, W. Lafayette IN (August 1981).

STECKE, KATHRYN E., "A Hierarchical Approach to Production Planning in Flexible Manufacturing Systems", Proceedings, Twentieth Annual Allerton Conference on Communication, Control, and Computing, Monticello IL (October 6-8, 1982).

STECKE, KATHRYN E., "Formulation and Solution of Nonlinear Integer Production Planning Problems for Flexible Manufacturing Systems", Management Science, Vol. 29, No. 3, pp. 273-288 (March 1983a).

STECKE, KATHRYN E., "On the Nonconcavity of Throughput in Certain Closed Queueing Networks", Working Paper No. 356, Division of Research, Graduate School of Business Administration, The University of Michigan, Ann Arbor MI (December 1983b).

STECKE, KATHRYN E. and MORIN, THOMAS L., "Optimality of Balancing in Flexible Manufacturing Systems", Working Paper No. 289, Division of Research, Graduate School of Business Administration, The University of Michigan, Ann Arbor MI (January 1982).

STECKE, KATHRYN E. and SCHMEISER, BRUCE W., "Alternative Representations of System Throughput in Closed Queueing Network Models of Multiserver Queues", Working Paper No. 324, Division of Research, The University of Michigan, Ann Arbor MI (December 1983).

STECKE, KATHRYN E. and SOLBERG, JAMES J., "The CMS Loading Problem", Report No. 20, NSF Grant No. APR 74 15256, School of Industrial Engineering, Purdue University, W. Lafayette IN (February 1981a).

STECKE, KATHRYN E. and SOLBERG, JAMES J., "Loading and Control Policies for a Flexible Manufacturing System", International Journal of Production Research, Vol. 19, No. 5, pp. 481-490 (September-October, 1981b).

STECKE, KATHRYN E. and SOLBERG, JAMES J., "The Optimality of Unbalanced Workloads and Machine Group Sizes for Flexible Manufacturing Systems", Working Paper No. 290, Division of Research, Graduate School of Business Administration, The University of Michigan, Ann Arbor MI (January 1982).

SURI, RAJAN, "Robustness of Queueing Network Formulas", Journal of the Association for Computing Machinery, Vol. 30, No. 3, pp. 564-594 (July 1983).