The optimality of balancing workloads in certain types of flexible manufacturing systems

Kathryn E. STECKE
Graduate School of Business Administration, University of Michigan, Ann Arbor, Michigan, U.S.A.

Thomas L. MORIN *
School of Engineering, Aristotlean University of Thessaloniki, Thessaloniki, Greece

Abstract. Symmetric mathematical programming is used to analyze the optimality of balancing workloads to maximize the expected production in a single-server closed queuing network model of a flexible manufacturing system (FMS). In particular, using generalized concavity we prove that, even though the production function is not concave, balancing workloads maximizes the expected production in certain types of m-machine FMS's with n parts in the system. Our results are compared and contrasted with previous models of production systems.

Keywords: Flexible manufacturing systems, workload balancing, symmetric mathematical programming, production planning, FMS loading

Introduction

Balance is basic to the human condition. Dividing equally is equitable in economics, democratic by definition, and just, according to Aristotle.[1] Balancing is also optimal in situations as diverse as maximizing the perimeters of inscribed polygons of a circle (Stark and Nicholls, 1972) and designing optimal binary search trees (Aho, Hopcroft, and Ullman, 1974 and Knuth, 1971). This paper considers balancing in the context of Computer Aided Manufacturing. Specifically, we use symmetric mathematical programming to establish the optimality of balanced workloads for certain types of flexible manufacturing systems.

A flexible manufacturing system (FMS) is an automated alternative to conventional means of batch manufacturing in the metal-cutting industry. An FMS consists of a number of numerically controlled machine tools which are linked together by an automated material handling system. Computers control most real-time activities such as the actual machining operations, part movements, and tool interchanges. An FMS can simultaneously and efficiently manufacture several part types. This combination of automation and increased flexibility offers the potential for vast improvements in productivity but as noted by

* On leave as a Fullbright Scholar from the School of Industrial Engineering, Purdue University, West Lafayette, Indiana.

[1] See p. 177 of Thomson's [1976] English translation or 1131 a11 of Bekker's [1831] Greek. Indeed Aristotle notes (p. 181, Thomson; 1132a30, Bekker) that the Greek word for just (*dikaion*) derives from equally dividing the whole in half (*dicha*).

We wish to acknowledge helpful discussions with Bruce W. Schmeiser, Leroy B. Schwarz, James J. Solberg and J. George Shanthikumar, as well as the useful comments of the referees. The research was supported in part by National Science Foundation Grant No. APR74 15256 to Purdue University.

Graves (1981), also increases the complexity of the problems faced by production managers. For example, the operation of an FMS requires a careful system set-up *prior to* production to achieve a good system utilization *during* production, even though technological hardware developments eliminate machine setup time. Several existing FMS's are described in Cavaillé, Forestier, and Bel (1981), Stecke and Solberg (1981b), Dupont-Gatelmand (1982), and Barash (1982).

This paper studies an idealized version of the *FMS loading problem*, which is one of the set-up problems of an FMS (see Stecke, 1983a). The loading problem involves determining the best allocation of operations and associated cutting tools of a set of part types among the machine tools subject to technological and capacity constraints.

The most widely applied loading objective is to balance, or equalize, the total workload assigned to each machine in: job shops (Deane and Moodie, 1972; Caie, Linden, and Maxwell, 1980); flow shops (Gutjahr and Nemhauser, 1964; Ignall, 1965; Magazine and Wee, 1979); and FMS's (Buzacott and Shanthikumar, 1980; Shanthikumar, 1982; Berrada and Stecke, 1983; Kusiak, 1983; Stecke and Talbot, 1983). However, the applicability and optimality of balancing has recently come under scrutiny. For example, Stecke and Solberg's (1981b) simulation results demonstrated that balancing workloads is not necessarily the best objective in an FMS. Other studies of finite-buffer stochastic flow lines also indicated that balancing the assigned workload is not always optimal (see Makino, 1964; Hillier and Boling, 1966, 1967; Payne, Slack, and Wild, 1972; Rao, 1976; Magazine and Silver, 1978; and El-Rayah, 1979). In particular, the numerical studies (see Hillier and Boling, 1966, 1967; and El-Rayah, 1979) discovered a 'bowl phenomenon' in which the expected production of a finite-buffered, balanced flow line is increased by assigning proportionately lower average processing or service times to the middle machines on the line.

Queueing network models have recently been used to analyze design issues and planning problems of FMS's. (For example, see Solberg, 1977; Buzacott and Shanthikumar, 1980; Cavaillé and Dubois, 1982; Dubois, 1983; Suri, 1983; and Stecke and Solberg, 1984.) Queueing networks have been shown to be robust models of FMS's even when the assumptions of the model are not satisfied (see Suri (1983) and Section 1 of this paper).

In the context of a closed queueing network (CQN), the loading problem is that of allocating a total amount of work among a system of (possibly grouped) machines so as to maximize expected production. Using a CQN, it has been shown that (Stecke and Solberg, 1984):

(i) the best way to partition the machine tools of a particular type into machine *groups* is to unbalance as much as possible the number of machines in each group;

(ii) for these better (unbalanced) system configurations, expected production is maximized by a particular unbalanced allocation of workload per machine.

However, in some practical situations, because of the discreteness of operation times, different machine tool requirements, and limited capacity tool magazines, balancing the workload per machine can be best even in some systems with grouped machines. This paper characterizes situations in which balancing is optimal: For those systems in which there is no grouping, or pooling machines of similar type into machine groups. Also, the fact remains that balancing is the almost universally applied loading objective, at least at present. Therefore, balancing *is* applicable to some FMS's.

In this paper we use a single-server CQN model to analyze the optimality of balancing for adequately buffered flexible manufacturing systems in which each operation is assigned to only one machine. We show that balancing maximizes the expected production in these systems. Specifically, symmetric mathematical programming and generalized concavity is used to establish the optimality of balanced workloads. The applications of these results are the algorithms to balance in FMS's. In particular, an efficient means of implementing a balancing FMS loading objective is provided in Berrada and Stecke (1983).

There is a related Computer Science literature. Price (1974), Trivedi and Kinicki (1978), Trivedi, Wagner, and Sigmon (1980), and Trivedi and Sigmon (1981) maximize throughput in central server, single class, single-server CQN subject to various cost constraints. The studies optimize different parameters such as service rate (of a CPU, say), capacity of servers (I/O devices), device speeds, and main memory size, subject to budgetary limitations. The parameters relate cost considerations to performance.

In this paper, a different non-central server CQN of single-server queues is considered. Rather than the budgetary constraints of the previous studies, we impose a constraint on total system workload that appears as a result of our unique scaling of workload and throughput. Therefore, the objective function and constraints are somewhat different. The motivation of our particular scaling results from our studies of optimal machine allocation and optimal workload assignment in FMS's.

Even though the objective function (to maximize expected production, or throughput) is not concave (see Stecke, 1983b), the production function is still well-behaved. In the situations studied here, the local maximum (which we prove is a balanced workload) is a global maximum.

The plan of the paper is as follows. The closed queueing network model is described in section 1. Notation and results from symmetric mathematical programming and generalized concavity that are required to characterize properties of optimal workloads are summarized in the Appendix. Properties of the production function and some preliminary results that are required to establish global optimality of balancing for this particular version of the *FMS loading problem* are provided in section 2. The main result is given in section 3. The paper concludes with a discussion of the relationships between this CQN and other models of manufacturing systems in section 4.

1. Closed queueing network model of an FMS

A flexible manufacturing system can be modeled as a closed network of arbitrarily-connected queues. The particular case of a central server CQN is depicted in Figure 1. There are m machines and n parts in the system. The average processing time of an operation by machine i is t_i, $i = 1, \ldots, m$.

Routing through the system is arbitrary, and can be described by the relative arrival rates (the q_i of Figure 1) to the machines. These can be obtained by any nonnegative solution to $q_i = \Sigma_j p_{ji} q_j$, where the p_{ij}'s are first-order Markovian probabilities. Our formulas permit any scaling of the q_i's. For example, if the q_i's are scaled to sum to one, q_i may be interpreted as the probability that a part leaving the load/unload station (L/UL) via a transporter goes next to machine i. Therefore, q_i is the expected number of visits to machine i per visit to the transporter (or L/UL). Other relevant routing possibilities are described in Stecke and Schmeiser (1983).

A measure of relative workload assigned to machine i is w_i (Buzen, 1973; Reiser and Kobayashi, 1975; Solberg, 1977), which is defined as the visit frequency times the average processing time, or $q_i t_i$, $i = 1, \ldots, m$. These workloads are relative since the q_i's can be scaled in any manner.

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

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

X_i is a scaled measure of workload, whose values lie between 0 and m, for all i.

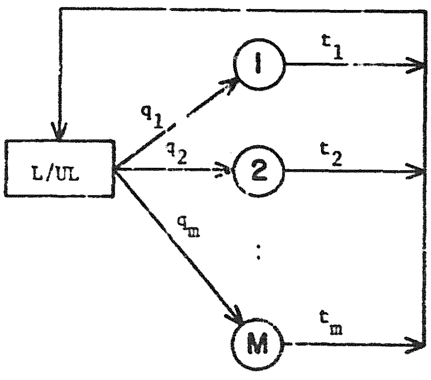

Figure 1. A closed queueing network model of a flexible manufacturing system.

There are several reasons for choosing this particular scaling: the total amount of work to be allocated among the machines is a fixed constant that equals the total number of machines: $\Sigma_{i=1}^{m} X_i = m$; the scaled workload is now independent of any particular, chosen scaling of the q_i; regardless of the number of machines, a balanced loading has a unit workload: $X_1 = X_2 = \ldots = X_m = 1$; this particular scaling of workload results in the production function (defined in equation (3) being a dimensionless function, whose values are also normalized to lie between zero and one; and finally, the workload scaling defined by (1) allows new alternative definitions of the production function (equation (3)). See Stecke (1981) or Stecke and Schmeiser (1983) for these alternative definitions. These are useful for providing insight into what this production function, associated with a CQN model, is.

A state of the CQN model of an FMS is given by $\tilde{n} = (n_1, \ldots, n_m)$, where n is the number of parts at machine tool i, both those waiting and those in process. For all i, $n_i \in \{0, 1, \ldots, n\}$ and $\Sigma_{i=1}^{m} n_i = n$. The steady-state probability of being in state $\tilde{n}$ is $p(\tilde{n}) = p(n_1, \ldots, n_m)$, which for this CQN model has the product form solution

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

where

$$G(m, n; X) = \sum_{n_1 + \ldots + n_m = n,\, n_i \geq 0} X_1^{n_1} X_2^{n_2} \ldots X_m^{n_m}, \quad i = 1, \ldots, m. \tag{2}$$

It can be seen that the function $G(m, n; X)$ is the normalizing constant that is required for the probabilities, $p(\tilde{n})$, to sum to 1. For the FMS depicted in Figure 1 with n parts in the system, the expected production rate, which is the expected number of parts produced per unit of time, can be defined as a function of $G(m, n; X)$, which in turn is a function of assigned workload, X_i. In fact, for a particular scaling of q_i, the production function, $\Pr(m, n; X)$, is given by Reiser and Kobayashi (1975) as

$$\Pr(m, n; X) = \frac{G(m, n-1; X)}{G(m, n; X)} = \frac{\sum_{n_1 + \ldots + n_m = n-1} X_1^{n_1} X_2^{n_2} \ldots X_m^{n_m}}{\sum_{n_1 + \ldots + n_m = n} X_1^{n_1} X_2^{n_2} \ldots X_m^{n_m}}. \tag{3}$$

The alternative definitions, referred to above, do provide additional intuition into just what $\Pr(m, n; X)$ means. For these insights, we refer the reader to Stecke and Schmeiser (1983).

As an example, the production function for two single machines and any number of parts, n, is

$$\begin{aligned}
\Pr(2, n; X) &= \frac{\sum_{n_1 + n_2 = n-1} X_1^{n_1} X_2^{n_2}}{\sum_{n_1 + n_2 = n} X_1^{n_1} X_2^{n_2}} \\
&= \frac{\sum_{n_1=0}^{n-1} X_1^{n_1} (2 - X_1)^{n-1-n_1}}{\sum_{n_1=0}^{n} X_1^{n_1} (2 - X_1)^{n-n_1}}, \quad \text{since } X_1 + X_2 = m = 2, \\
&= \frac{X_1^n - (2 - X_1)^n}{X_1^{n+1} - (2 - X_1)^{n+1}},
\end{aligned} \tag{4}$$

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

Many performance measures that can be obtained from CQN models, such as the expected production rate, are insensitive to the form of the service time distribution–see Helm and Schassberger (1982) and

Dukhovny and Koenigsberg (1981). In fact, for the performance measure of expected production, the service time distribution can be arbitrary.

The assumptions of our CQN model of a flexible manufacturing system are that:

1. There are n parts (or pallets) circulating through a system of m machines.

2. There is a buffer at each machine tool that has the capacity to hold all n parts, including the part being machined.

3. The queue discipline at each machine tool can be either FCFS, infinite server, LCFS preempt-resume, processor sharing (see Baskett et al., 1975), random selection, or one developed by Kelly (1979) which allows an arbitrary distribution to be defined at each node.

The main restrictive assumption is the limited number of allowable queue disciplines, which is why product form queueing networks are not used to study scheduling problems.

Queueing network models have been shown to be accurate in qualitatively predicting steady-state behavior of FMS's. For example, Solberg (1977) compared results from his CQN computer program, CAN-Q, to those of a detailed simulation of the Sundstrand/Caterpillar FMS of Peoria, Illinois (Stecke, 1977) to find that the performance measures of all machine utilizations and expected production rate differed from those of the FMS by less than 3 percent. Similar results were observed by Kimemia and Gershwin (1978), Secco-Suardo (1978), and Dubois (1983). Queueing network models have also been used to model other nonmanufacturing systems in which the service time distributions were not exponential, with encouraging results. For example, Hughes and Moe (1973), Giammo (1976), Lipsky and Church (1977), and Rose (1976, 1978) have verified in empirical studies that queueing network models reproduce observed quantities with reasonable accuracy. Attempts to explain the observed robustness through operational analysis can be found in Denning and Buzen (1978) and Suri (1983).

2. Preliminary results

The production function given in equation (3) is difficult to characterize analytically. However, it can be evaluated numerically using Buzen's (1973) efficient algorithm. The function behaves so well empirically that some researchers (i.e., Secco-Suardo, 1978; and Solberg, 1979) have conjectured that it must be concave. Concavity would be desirable because it would insure that a local maximum, if it exists, is a global maximum. However, Stecke (1983b) has shown that, contrary to conjecture, the production function is *not* concave in general, even though it is concave in a few restrictive cases. Fortunately, however, the function satisfies weaker generalized concavity conditions, which are also sufficient to insure that a local maximum is a global maximum.

Using the Definitions (**D**) and Theorems (**T**) in the Appendix, we first establish two preliminary results on symmetric mathematical programming which are used subsequently in section 3 to prove the main result.

Proposition 1. *The set χ of feasible loadings is a closed S-convex set.*

Proof. From (1), we have

$$\chi = \left\{ (X_1, \ldots, X_m) \mid \sum_{i=1}^{m} X_i = m,\ X_i \in \boldsymbol{R},\ 0 \leqslant X_i \leqslant m \right\}.$$

Therefore, χ is clearly closed. It is also clearly convex and symmetric. Then by **T11**, χ is S-convex. □

Proposition 2. *The quotient of two symmetric functions in the same variables on the same symmetric set χ is a symmetric function.*

Proof. Suppose that $f(x)$ and $g(x)$ are symmetric functions on the symmetric set χ. Then by **D5**,

$$f(xP)=f(x) \quad \text{and} \quad g(xP)=g(x)$$

for all $x \in \chi$ and for any permutation matrix P. Let $h(x)=f(x)/g(x)$ for all $x \in \chi$ such that $g(x) \neq 0$. Then

$$h(xP)=f(xP)/g(xP)=f(x)/g(x)=h(x).$$

Therefore, $h(x)$ is a symmetric function on χ. □

Proposition 2 will be used subsequently in Lemma 4 to prove that the production function is symmetric. Prior to doing that, we prove directly the S-concavity of the production function for two machines, $\Pr(2, n; X)$, in Theorem 3 that follows.

Theorem 3. $\Pr(2, n; X)$ *is S-concave.*

Proof. From (4), we have

$$\begin{aligned}\Pr(2, n; X) &= \frac{X_1^n-(2-X_1)^n}{X_1^{n+1}-(2-X_1)^{n+1}}\\ &= \frac{X_1^n - X_2^n}{X_1^{n+1} - X_2^{n+1}}, \quad \text{for } X_1 \neq X_2;\\ &= n/(n+m-1), \quad \text{for } X_1 = X_2.\end{aligned}$$

Differentiating yields:

$$\frac{\partial \Pr(2, n; X)}{\partial X_2} = \frac{-(X_1^{n+1}-X_2^{n+1})nX_2^{n-1}+(X_1^n-X_2^n)(n+1)X_2^n}{(X_1^{n+1}-X_2^{n+1})^2}.$$

Therefore,

$$\begin{aligned}(X_2-X_1)&\left(\frac{\partial \Pr(2, n; X)}{\partial X_2} - \frac{\partial \Pr(2, n; X)}{\partial X_1}\right)\\ &= \{(X_2-X_1)\{[(n+1)X_2^n(X_1^n-X_2^n)-nX_2^{n-1}(X_1^{n+1}\\ &\quad -[nX_1^{n-1}(X_1^{n+1}-X_2^{n+1})-(n+1)X_1^n(X_1^n-X_2^n)]\}\end{aligned}$$

Since the denominator is positive for $X_1 \neq X_2$, it may be dropped, yielding after simplification:

$$\begin{aligned}&(X_2-X_1)\{X_2^{n-1}[(n+1)X_2(X_1^n-X_2^n)-n(X_1^{n+1}-X_2^{n+1})]\\ &\quad +X_1^{n-1}[(n+1)X_1(X_1^n-X_2^n)-n(X_1^{n+1}-X_2^{n+1})]\},\end{aligned}$$

which upon rearranging

$$=(X_2-X_1)(X_1^2-X_2^2)\left\{\sum_{i=1}^{n-1} X_1^{2n-2-2i}X_2^{2i} - nX_1^{n-1}X_2^{n-1}\right\}. \tag{5}$$

In order to show that (5) is not positive, it suffices to show that

$$X_1^{2n-2-2i}X_2^{2i}+X_1^{2i}X_2^{2n-2-2i} \geqslant 2X_1^{n-1}X_2^{n-1}, \tag{6}$$

since the summation of (5) can be separated into $n/2$ inequalities of the form (6). Assume that $2i < 2n-2$.

Subtracting the RHS from the LHS of (6) yields:

$$\begin{aligned} &X_1^{2i}X_2^{2i}\left(X_1^{2n-2-4i}-2X_1^{n-1-2i}X_2^{n-1-2i}+X_2^{2n-2-4i}\right)\\ &\quad= X_1^{2i}X_2^{2i}\left(X_1^{2(n-1-2i)}-2X_1^{n-1-2i}X_2^{n-1-2i}+X_2^{2(n-1-2i)}\right)\\ &\quad= X_1^{2i}X_2^{2i}\left(X_1^{n-1-2i}-X_2^{n-1-2i}\right)^2\\ &\quad\geqslant 0. \end{aligned}$$

Therefore, (6) holds for all $i < n-1$. The proof for $2i > 2n-2$ follows *mutatis mutandis* and equality holds if $i = n-1$. □

Next consider the m machine case.

Lemma 4. $\Pr(m, n; X)$ is symmetric.

Proof. By **D5** and **T9**, $\Pr(m, n; X)$ is symmetric if the value of the function remains the same when the X_i are permuted.

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

$G(m, n; X)$ is symmetric. Since the production function is the quotient of two symmetric functions, by Proposition 2, $\Pr(m, n; X)$ is a symmetric function of X. □

If in addition $\Pr(m, n; X)$ is quasiconcave, then by **T12** $\Pr(m, n; X)$ is S-concave and we can use **T15** to prove that balancing is optimal.

Theorem 5. *The production function*, $\Pr(m, n; X)$, *is strictly quasiconcave.*

Proof. The result is provided in Stecke (1983b). □

3. Characterizing optimal workloads

We now state and prove the main result.

Theorem 6. *A balanced allocation of workload maximizes expected production, i.e.*,

$$X^* = [X_1, X_2, \ldots, X_m] = [1, 1, \ldots, 1].$$

Proof. By Proposition 1, the set of feasible loadings, χ, is closed and S-convex.
By Lemma 4, $\Pr(m, n; x)$ is symmetric. By **T12**, since $\Pr(m, n; X)$ is quasiconcave by Theorem 5, then $\Pr(m, n; X)$ is S-concave. By **T15**, the set χ^* of points maximizing $\Pr(m, n; X)$ over the set χ is a closed S-convex set. χ^* is not empty since $\Pr(m, n; X) \in [0, 1]$ for all m, n, and $X \in [0, m]$. The symmetric point of χ is the point [1, 1, ...,]. By **T14**, $[1, \ldots 1] \in \chi^*$.
Therefore, a balanced allocation maximizes the expected production. □

Balancing is now justified for the systems examined here, i.e., FMS's with no pooling of similar machines.

We next provide some numerical results. Specifically, the following computer-drawn graphs demonstrate the behavior of the production function. First, Figure 2 is a graph of $\Pr(2, n; X)$ as a function of X_1 for $n = 4, 5, \ldots, 14$ and infinity. For each curve, 400 points (X, $\Pr(2, n; X)$) were plotted. These were calculated using a variation of Solberg's CAN-Q program [1980]. The maximum functional value, also

Table 1
Maximum (balanced) production rates and corresponding workloads for two-machine systems

n	t_1	t_2	X_1	X_2	Maximum (balanced) production rate
4	1.0	1.0	1.0	1.0	0.800
5	1.0	1.0	1.0	1.0	0.833
6	1.0	1.0	1.0	1.0	0.857
7	1.0	1.0	1.0	1.0	0.875
8	1.0	1.0	1.0	1.0	0.889
9	1.0	1.0	1.0	1.0	0.900
10	1.0	1.0	1.0	1.0	0.909
11	1.0	1.0	1.0	1.0	0.917
12	1.0	1.0	1.0	1.0	0.923
13	1.0	1.0	1.0	1.0	0.929
14	1.0	1.0	1.0	1.0	0.933
∞	1.0	1.0	1.0	1.0	1.000

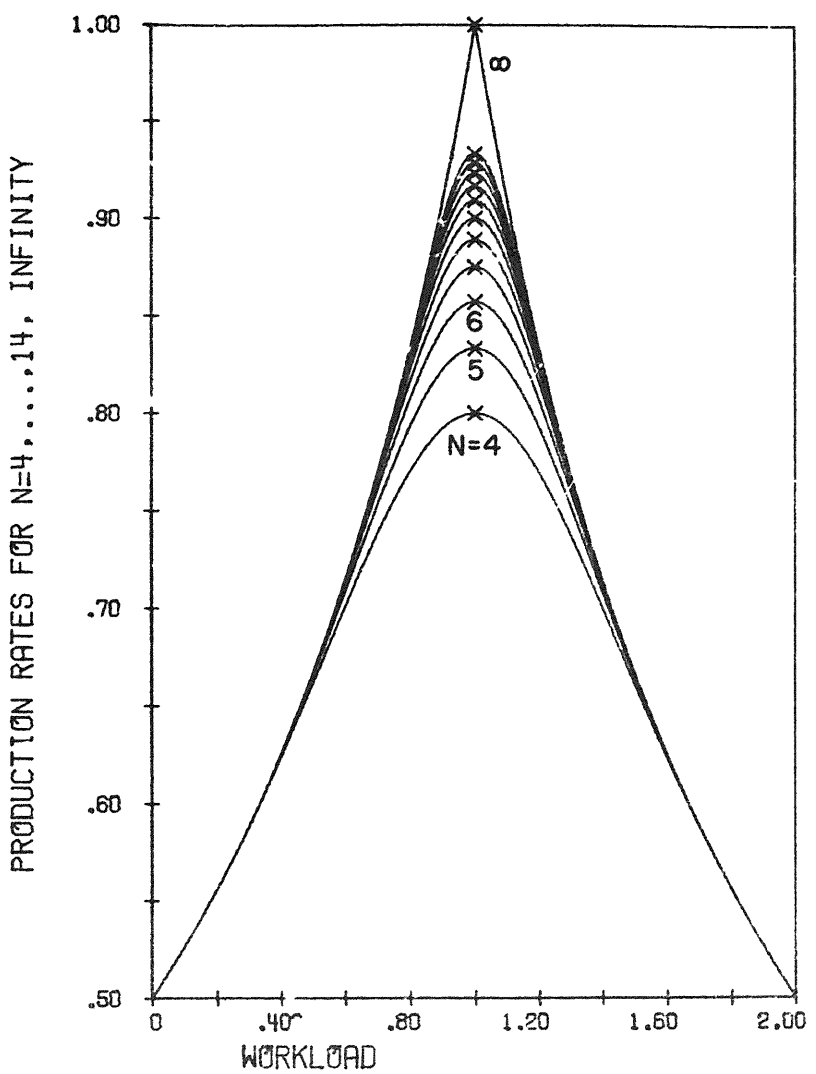

Figure 2. Production rate as a function of workload assigned to machine 1 for 2-machine systems.

Table 2
Maximum (balanced) production rates and corresponding workloads for three-machine systems

n	Balanced production rate	Maximum production rate	$X_1^* = t_1$	$X_2^* = t_2$
4	0.667	0.667	1.0	1.0
5	0.714	0.714	1.0	1.0
6	0.750	0.750	1.0	1.0
7	0.778	0.778	1.0	1.0
8	0.800	0.800	1.0	1.0
9	0.818	0.818	1.0	1.0
10	0.833	0.833	1.0	1.0
11	0.846	0.846	1.0	1.0
12	0.857	0.857	1.0	1.0
13	0.867	0.867	1.0	1.0
14	0.875	0.875	1.0	1.0
∞	1.000	1.000	1.0	1.0

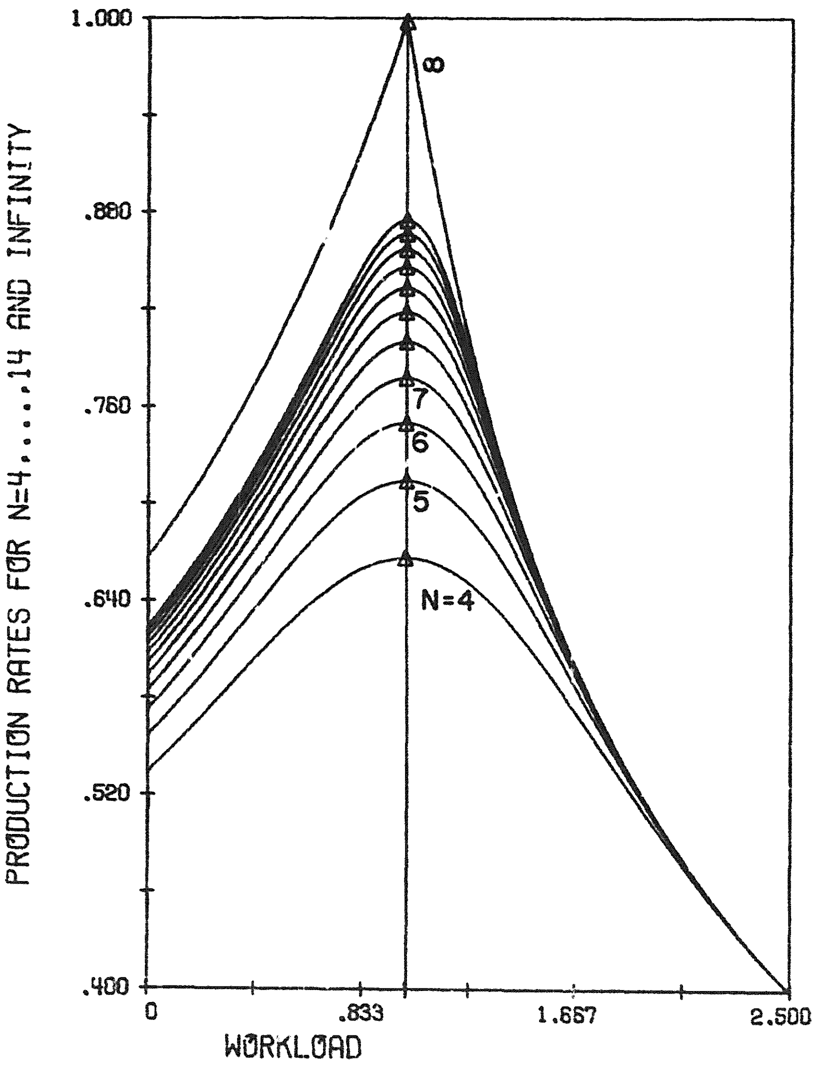

Figure 3. Production rate as a function of workload assigned to machine 1 for 3-machine systems.

calculated and plotted, was attained at $X = [1,1]$. Table 1 displays the calculated optimal allocation ratios and the maximum normalized production rate for each n. For $q_1 = q_2 = 0.5$ (each machine is visited half of the time on the average), the average processing times, t_1 and t_2, vary so that $t_1 + t_2 = 2$. The optimal allocation occurs when $t_1 = t_2 = 1$. Then $X_i = 2q_i t_i/(q_1 t_1 + q_2 t_2) = t_i$, where i is 1 or 2. The optimal allocation of workload in this system is balanced.

Figure 3 is a graph of Pr(3, n; X) as a function of X_1 for $n = 4, 5, \ldots, 14$ and infinity, along the plane $X_2 = X_3$. It is interesting to note that this two-dimensional slice of Pr(3, n; X) is nonsymmetric even though the entire function is symmetric. The maximum is shown to be at $X_1 = X_2 = X_3 = 1$. The computer program generated both the balanced and the maximum normalized productions, as well as the optimal allocation ratios. These are shown for each n in Table 2.

Finally, Figure 4 displays Pr(4, n; X) for $n = 4, \ldots, 14$ and infinity along the intersection of the planes $X_1 = X_2$ and $X_3 = X_4$. Table 3 gives values for Pr(4, n; X). Notice that for all finite n, Pr(3, n; X) > Pr(4, $n + 1$; X). That is, as the number of machines increases, the actual expected production obviously increases but the normalized expected production decreases. The apparent anomaly is the result of the normalization of production to the scaling between 0 and 1.

We conclude that even though Pr(m, n; X) is *not* concave for any $m \geq 2$ and $n > 2$, balancing *is* optimal for all cases (fixed-route FMS's) considered here.

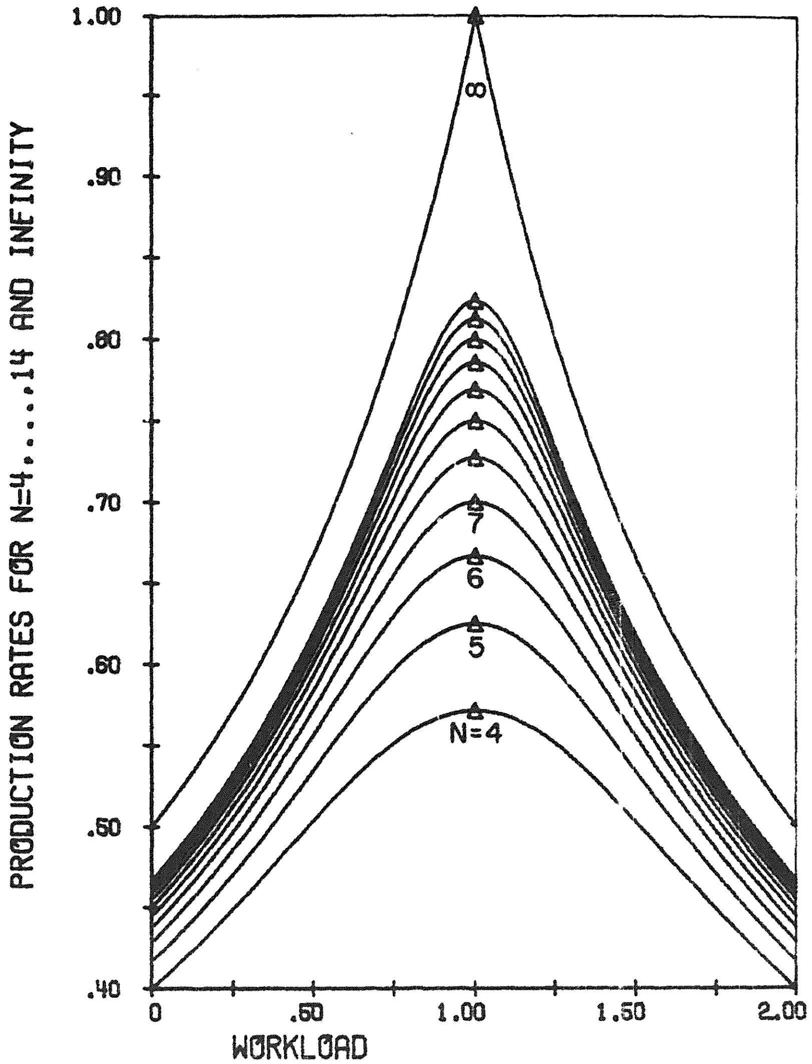

Figure 4. Production rate as a function of workload assigned to machine 1 for 4-machine systems.

Table 3
Maximum (balanced) production rates and corresponding workloads for four-machine systems

n	Balanced production rate	Maximum production rate	$X_1^* = t_1$	$X_2^* = t_2$
4	0.571	0.571	1.0	1.0
5	0.625	0.625	1.0	1.0
6	0.667	0.667	1.0	1.0
7	0.700	0.700	1.0	1.0
8	0.727	0.727	1.0	1.0
9	0.750	0.750	1.0	1.0
10	0.769	0.769	1.0	1.0
11	0.786	0.786	1.0	1.0
12	0.800	0.800	1.0	1.0
13	0.812	0.812	1.0	1.0
14	0.824	0.824	1.0	1.0
∞	1.000	1.000	1.0	1.0

4. Discussion

The results can be related to similar studies of workload allocation in manufacturing systems. Our results differ from the finite-buffer stochastic flow shop studies (Hillier and Boling, 1966, 1967; Magazine and Silver, 1978) mainly because we assume an adequate buffer at each machine.

In fact, using our CQN model, the expected production is identical for both flow shops and job shops in which each operation is assigned to only one machine. To see this, let t_i (the average processing time of an operation by machine i) be identical for both systems. The routing mechanism, defined by Markovian probabilities p_{ij}, for a three-machine flow shop is given by the following transition matrix:

$$P_F = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix};$$

the routing for the job shop is given by:

$$P_J = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

Then solving balance equations

$$q_{i(k)} = \sum_{j=1}^{m} p_{ij(k)} q_{(k)}, \quad i = 1, \ldots, m, \ (k) = F \text{ or } J,$$

for the two systems produces identical, steady state results. That is,

$$q_{i(F)} = q_{i(J)}, \quad i = 1, \ldots, m.$$

Since q_i, t_i, m, and n are identical for both the flow and job shops, the expected production rate is the same for both systems.

Some implications are as follows. Under the assumptions of our CQN model (in particular, random processing times and adequate buffer at each machine), the two extreme system types (flow and job shops) are equally efficient. Intuition indicates that as variability decreases, the flow shop becomes more efficient than the job shop.

The Hillier and Boling (1966, 1967) result is basically the following:

For a stochastic flow shop with a *finite* buffer at each machine, the expected production is maximized by a specific *unbalancing* in the workload assigned to each machine.

Our analogous result is:

For a stochastic flow shop with an *infinite* buffer at each machine, the expected production is maximized by *balancing* the workload assigned to each machine.

In other words, as buffer size increases, the degree of unbalance in the optimal workload decreases, until in the limit, a balanced schedule is optimal.

Appendix. Results from generalized concavity and symmetric mathematical programming

Definitions and previously published results which are required to prove the optimality of balanced workloads are reviewed below. The definitions and results concerning generalized concavity can be found in Mangaserian (1969) or Bazaraa and Shetty (1979), those concerning S-concavity can be found in Berge (1963), and those on symmetric mathematical programming can be found in Greenberg and Pierskalla (1970).

Let f be a real-valued function mapping $\chi \to \boldsymbol{R}$, where χ is a closed subset of $\boldsymbol{R}^m$. We require the following Definitions (D) and Theorems (T):

D1. f is a *quasiconcave* function on the nonempty convex set $\chi \subseteq \boldsymbol{R}^m$ if and only if (iff) for any two points $x^1, x^2 \in \chi$, and for all $\lambda \in [0,1]$,

$$f(\lambda x^1 + (1-\lambda)x^2) \geqslant \min\{f(x^1), f(x^2)\}.$$

D1 is not enough to insure that a local maximum is a global maximum. For this to be true we have:

D2. f is a *strictly quasiconcave* function on the convex set $\chi \subseteq \boldsymbol{R}^m$ iff for any two points $x^1, x^2 \in \chi$, and for all $\lambda \in (0, 1)$, with $f(x^1) \neq f(x^2)$,

$$f(\lambda x^1 + (1-\lambda)x^2) > \min\{f(x^1), f(x^2)\}.$$

In order to insure that a global maximum is unique, we require:

D3. f is a *strongly quasiconcave* function on the convex set $\chi \subseteq \boldsymbol{R}^m$ iff for any two points $x^1, x^2 \in \chi$ such that $x^1 \neq x^2$ and for all $\lambda \in (0, 1)$,

$$f(\lambda x^1 + (1-\lambda)x^2) > \min\{f(x^1), f(x^2)\}.$$

D4. χ is a *symmetric set* if $x \in \chi \Rightarrow x\boldsymbol{P} \in \chi$ for all permutation matrices $\boldsymbol{P}$, where $\boldsymbol{P}$ is a *permutation matrix* if

(i) each row has only one entry equal to one;
(ii) each column has only one entry equal to one; and
(iii) all remaining entries are equal to zero.

D5. f is a *symmetric function* on a symmetric set χ if for any permutation matrix $\boldsymbol{P}$

$$f(x\boldsymbol{P}) = f(x) \quad \text{for all } x \in \chi.$$

D6. χ is *S-convex* if $x \in \chi \Rightarrow xS \in \chi$ for all doubly stochastic matrices S, where S is a *double stochastic matrix* of order m if all of its entries, p_{ij}, satisfy

(i) $p_{ij} \geqslant 0$ for all i, j;

(ii) $\sum_{i=1}^{m} p_{ij} = 1$ for all j,

(iii) $\sum_{j=1}^{m} p_{ij} = 1$ for all i.

D7. f is a *(strictly) S-concave function* on an S-convex set χ if for any S

$$f(xS)(>) \geqslant f(x) \quad \text{for all } x \in \chi.$$

T8. *Let D be an open interval in $\mathbf{R}$ and let f be a symmetric differentiable function in $D^m \subseteq \mathbf{R}^m$. If for all $x = (x_1, \ldots, x_m) \in D^m$ such that $x_1 \neq x_2$ we have*

$$(x_2 - x_1)\left(\frac{\partial f}{\partial x_2} - \frac{\partial f}{\partial x_1}\right)(<) \leqslant 0,$$

then the function is (strictly) S-concave in D^m (Berge, 1963, Theorem 5, p. 221).

T9. *An S-concave function f in $\mathbf{R}^m$ is symmetric in the components $x_1, \ldots, x_m$* of $x \in \chi$; *that is, the value of $f(x_1, \ldots, x_m)$ remains the same when the x_i are permuted* (Berge, 1963, Theorem 3, p. 220).

T10. *If D is an open interval in $\mathbf{R}$, a necessary and sufficient condition for a differentiable and symmetric function f to be (strictly) S-concave in D^m is that for all $x_1, x_2 \in D$,*

$$(x_2 - x_1)\left(\frac{\partial f}{\partial x_2} - \frac{\partial f}{\partial x_1}\right)(<) \leqslant 0$$

(Berge, 1963, Theorem 6, p. 224).

T11. *Symmetric convex sets are S-convex (but not necessarily conversely).*

T12. *Symmetric (strictly) quasiconcave functions defined on a symmetric convex set χ are (strictly) S-concave (but not necessarily conversely).*

D13. A *point* $x = (x_1, x_2, \ldots, x_m)$ is *symmetric* iff $x_i = y$, $\forall\, i = 1, \ldots, m$.

T14. *Every nonempty S-convex set contains a symmetric point.*

T15. *If χ is a closed, S-convex set and f is S-concave on χ, then the set χ° of points maximizing f over χ is a closed S-convex set* (Greenberg and Pierskalla, 1970).

References

Aho, A.V., Hopcroft, J.E. and Ullman, J.D. (1974), *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading.

Aristotle, *Nicomachean Ethics*, Book V, Justice (1976), pp. 171–202, translated by J.A.K. Thomson, rev. H. Tredennick, Penguin Books Ltd., Harmondsworth, Middlesex, England.

Barash, Moshe M. (1982), "Computerized manufacturing systems for discrete products," Ch. VII-9 in Salvendy, G. (ed.), *The Handbook of Industrial Engineering*, John Wiley & Sons, New York.

Baskett, Forest, K. Mani Chandy, Muntz, R. and Palacios, F.G. (1980), "Open, closed, and mixed networks with different classes of customers," *Journal of the Association for Computing Machinery* 22 (2), 248–260.

Bazaraa, M. S., and Shetty, C.M. (1979), *Nonlinear Programming Theory and Algorithms*, John Wiley & Sons, New York.

Bekker, I. (1831), (ed.), *Aristotelis Opera*, Berlin.

Berge, C. (1963), *Topological Spaces*, The Macmillan Company, New York.

Berrada, M., and Stecke, K.E. (1983), "A branch and bound approach for machine loading in flexible manufacturing systems," Working Paper No. 329, Division of Research, Graduate School of Business, University of Michigan, Ann Arbor, MI.

Buzacott, J.A., and Shanthikumar, J.G., (1980), "Models for understanding flexible manufacturing systems," *AIIE Transactions* 12 (4) 339–350.

Buzen, J.P. (1973), "Computational algorithms for closed queueing networks with exponential servers," *Communications of the Association for Computing Machinery* 16 (9), 527–531.

Caie, J., Linden, J. and Maxwell, W.L., (1980), "Solution of a single stage machine load planning problem," *Omega* 8 (3), 355–360.

Cavaillé, J.-B., and Dubois, D. (1982), "Heuristic methods based on mean-value analysis for flexible manufacturing systems performance evaluation," *Proceedings of the 21st IEEE Conference on Decision and Control*, Orlando, 1061–1065.

Cavaillé, J.-B. Forestier, J.P. and Bel, G., (1981), "A simulation program for analysis and design of a flexible manufacturing system," *Proceedings of the IEEE Conference on Cybernetics and Society*, Atlanta, 257–259.

Deane, R.H., and Moodie, C.L., (1972), "A dispatching methodology for balancing workload assignments in a job shop production facility," *AIIE Transactions* 4, 277–283.

Denning, P.J., and Buzen, J.P. (1978), "The operational analysis of queueing network models," *Computing Surveys*, 10 (3), 225–261.

Dubois, D. (1983), "A mathematical model of a flexible manufacturing system with limited in-process inventory," *European Journal of Operational Research* 14 (1), 66–78.

Dupont-Gatelmand, C. (1982), "A survey of flexible manufacturing systems," *Journal of Manufacturing Systems* 1 (1), 1–16.

Dukhovny, I.M., and Koenigsberg, E., (1981). "Invariance properties of queueing networks and their application to computer/communications systems," *Information Systems and Operations Research* 19 (3), 185–204.

El-Rayah, T.E. (1979), "The efficiency of balanced and unbalanced production lines," *International Journal of Production Research* 17 (1), 61–75.

Giammo, T. (1976), "Validation of a computer performance model of the exponential queueing network family, "*Proceedings of the International Symposium of Computer Performance Modeling, Measurement, and Evaluation*, Harvard University. Also published in *Acta Informatica* 17 (2), 137–152.

Graves, S.C. (1981), "A review of production scheduling," *Operations Research* 29, 646–675.

Greenberg, H.J., and Pierskalla, W.P. (1970), "Symmetric mathematical programs," *Management Science* 16 (5), 309–312.

Gutjahr, A.L., and Nemhauser, G.L. (1964), "An algorithm for the line balancing problem," *Management Science* 11, 308–315.

Helm, W.E., and Schassberger, R. (1982), "Insensitive generalized semi-Markov schemes with point process input," *Mathematics of Operation Research* 7(1), 129–138.

Hillier, F.S., and Boling, R.W. (1966), "The effect of some design factors on the efficiency of production lines with variable operation times," *Journal of Industrial Engineering* 17 (5), 657–658.

Hillier, F.S., and Boling, R.W. (1967), "Finite queues in series with exponential or Erlang service times: A numerical approach," *Operations Research* 15 (2), 286–303.

Hughes, P.H., and Moe, G. (1973), "A structural approach to computer performance analysis," *Proceedings of the National Computer Conference*, AFIPS Press, Montvale 42, 109–119.

Ignall, E.J. (1965), "A review of assembly line balancing," *Journal of Industrial Engineering* 16 (4), 43–52.

Kelly, F.P., (1979), *Reversibility and Stochastic Networks*, John Wiley & Sons, New York.

Kimemia, J., and Gershwin, S.B., (1978), "Multicommodity network flow optimization in flexible manufacturing systems," Report No. ESL-FR-834-2, M.I.T., Cambridge.

Knuth, D.E. (1971), "Optimal binary search trees," *Acta Informatica* 1, 14–25.

Kusiak, A., (1983), "Loading models in flexible manufacturing systems," *Proceedings of the Seventh International Conference on Production Research*, Windsor, Ontario.

Lipsky, L., and Church J.D., (1977), "Applications of a queueing network model for a computer system," *Computing Surveys* 9, 205–221.

Magazine, M.J., and Wee, T.S., (1979), "The generalization of bin-packing heuristics to the line balancing problem," Working Paper No. 128, Department of Management Sciences, University of Waterloo, Ontario.

Magazine, M.J., and Silver, G.L., (1978), "Heuristics for determining output and work allocations in series flow lines.," *International Journal of Production Research* 16 (6), 169–181.

Makino, T., (1964), "On the mean passage time concerning some queueing problems of the tandem type," *Journal of the Operations Research Society of Japan* 7, 17–47.

Mangaserian, O.L. (1969), *Nonlinear Programming*, McGraw-Hill Company, New York.

Payne, S., Slack, N. and Wild, R., (1972), "A note on the operating characteristics of balanced and unbalanced production flow lines," *International Journal of Production Research* 10 (1) 93–98.

Price, Gordan, T., (1974), "Probability models of multiprogrammed computer systems," Ph.D. dissertation, Department of Electrical Engineering, Stanford University, Stanford.

Rao, N.P. (1976), "A generalization of the 'Bowl Phenomenon' in series production systems," *International Journal of Production Research*, 14, 437–443.

Reiser, M., and Kobayashi H., (1975), "Horner's role for the evaluation of general closed queueing networks," *Communications of the Association for Computing Machinery* 18 (10), 592–593.

Rose, C.A., (1976), "Validation of a queueing model with classes of customers," *Proceedings of the International Symposium on Computer Performance Modeling, Measurement, and Evaluation,* Harvard University, Cambridge, 318–325.

Rose, C.A. (1978), "A measurement procedure for queueing network models of computer systems," *Computing Surveys*, 10, 263–280.

Secco-Suardo, G., "Optimization of a closed network of queues," Report No. ESL-FR-834-3, Electronic Systems Laboratory, M.I.T., Cambridge.

Shanthikumar, J.G. (1982), "On the superiority of balanced load in a flexible manufacturing system", Technical report.

Solberg, J.J. (1977), "A mathematical model of computerized manufacturing systems," *Proceedings of the International Conference on Production Research,* Tokyo.

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.

Solberg, J.J. (1980), "CAN-Q user's guide," Report No. 9 (Revised), NSF Grant No. APR74 15256, School of Industrial Engineering, Purdue University, West Lafayette.

Stark, R.M., and Nicholls, R.L., (1972), *Mathematical Foundation for Design,* McGraw-Hill, New York.

Stecke, K.E., (1977), "Experimental investigation of a computerized manufacturing system," Master's Thesis, School of Industrial Engineering, Purdue University, West Lafayette.

Stecke, K.E. (1981), "Production planning problems for flexible manufacturing systems," Ph.D. dissertation, Purdue University, West Lafayette.

Stecke, K.E. (1982), "A hierarchical approach to production planning in flexible manufacturing systems," *Proceedings of the Twentieth Annual Allerton Conference on Communication, Control, and Computing,* Monticello.

Stecke, K.E., (1983a), "Formulation and solution of nonlinear integer production planning problems for flexible manufacturing systems," *Management Science* 29 (3), 273–288.

Stecke, K.E. (1983b), "On the nonconcavity of throughput in certain closed queueing networks," Working Paper No. 356, Division of Research, Graduate School of Business Administration, University of Michigan, Ann Arbor.

Stecke, K.E., and B.W. Schmeiser (1983), "Alternative representations of system throughput in closed queueing network models of multiserver queues," Working Paper No. 324, Division of Research, Graduate School of Business Administration, The University of Michigan, Ann Arbor.

Stecke, K.E., and Solberg, J.J., (1981a), "The CMS loading problem," Report No. 20, NSF Grant No. APR 74 15256, School of Industrial Engineering, Purdue University, West Lafayette.

Stecke, K.E., and Solberg, J.J., (1981b), "Loading and control policies for a flexible manufacturing system," *International Journal of Production Research* 19 (5) 481-490.

Stecke, K.E. and Solberg J.J., (1984), "The optimality of unbalancing both workloads and machine group sizes in closed queueing networks of multiserver queues," *Operations Research*, forthcoming.

Stecke, K.E., and Talbot, F.B., (1983), "Heuristic loading algorithms for flexible manufacturing systems," *Proceedings of the Seventh International Conference on Production Research*, Windsor, Ontario.

Suri, R., (1983), "Robustness of queueing network formulas," *Journal of the Association for Computing Machinery* 30 (3), 564–594.

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

Trivedi, K.S., and Sigmon, T.M., (1981), "Optimal design of linear storage hierarchies," *Journal of the Association for Computing Machinery* 28 (2), 270–288.

Trivedi, K.S., Wagner, R.A. and Sigmon, T.M., (1980), "Optimal selection of CPU speed, device capabilities and file assignments," *Journal of the Association for Computing Machinery* 27 (3), 457–473.