OPTIMALITY OF BALANCED WORKLOADS IN FLEXIBLE MANUFACTURING SYSTEMS

Working Paper No. 289-c

Kathryn Stecke
Graduate School of Business Administration
The University of Michigan, Ann Arbor Michigan

and

Thomas L. Morin
School of Industrial Engineering
Purdue University, West Lafayette Indiana

European Journal of Operational Research, forthcoming
ABSTRACT

Generalized concavity and symmetric mathematical programming are used to analyze the optimality of balanced workloads to maximize the expected production of flexible manufacturing systems (FMSs) using a single-server closed queueing network model. In particular, we prove that balancing workloads maximizes expected production in certain types of $m$-machine FMSs with $n$ parts in the system. Our results are compared and contrasted with previous models of production systems.
A flexible manufacturing system (FMS) is an automated alternative to the conventional means of batch manufacturing in the metal-cutting industry that 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 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 set-up time. Several existing FMSs are described in Cavaille, 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 FMSs (Buzacott and Shanthikumar [1980], 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 produc-
tion of a finite-buffered, balanced flow line is increased by assigning pro-
portionately 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 FMSs. (For example, see Buzacott and Shanthikumar
[1980], Cavaillé and Dubois [1982], Dubois [1983], Stecke and Solberg [1982],
and Suri [1983].) Queueing networks have been shown to be robust models of
FMSs even when the assumptions of the model are not satisfied (see Suri [1983]
and §1).

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 [1982]):

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 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. 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. Balancing is applicable to some FMSs.

In particular, 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, we use generalized concavity and symmetric mathematical programming to establish the optimality of balanced workloads. 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 FMSs.
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 §1. Notation and results from generalized concavity and symmetric mathematical programming 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 §2. The main result is given in §3. A discussion of the relationships between this CQN and other models of manufacturing systems is contained in §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 \( t_i \), \( i = 1, \ldots, m \).

![Figure 1](image)

A Closed Queueing Network Model of a Flexible Manufacturing System.
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 \( q_i = \sum_j p_{ij} q_j \) 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 \( \sum_{j=1}^{m} q_j t_j / m \) is the average workload per machine, to provide:

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

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

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: \( \sum_{i=1}^{m} X_i = m \); 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) below) being a dimensionless function, whose values are also normalized to lie between zero and one; and finally, the workload
scaling defined by equation (1) allows new alternative definitions of the production function (equation (3) below). 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 \( n_i \) 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 \( \sum_{i=1}^{m} n_i = n \). 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: \( p(\tilde{n}) = \frac{1}{G(m,n;X)} \frac{n_1^{n_1} n_2^{n_2} \cdots n_m^{n_m}}{X_1^{n_1} X_2^{n_2} \cdots X_m^{n_m}} \), where:

\[
G(m,n;X) = \sum_{n_1 + \ldots + n_m = n} \frac{n_1^{n_1} n_2^{n_2} \cdots n_m^{n_m}}{X_1^{n_1} X_2^{n_2} \cdots X_m^{n_m}}, \quad i=1, \ldots, m. \tag{2}
\]

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, \( X_1 \).

In fact, for a particular scaling of \( q_i \), the production function, \( \Pr(m,n;X) \), is given by Reiser and Kobayashi [1975] as:

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

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:
\[ \Pr(2,n;X) = \frac{\sum_{x_1+x_2=m-1}^{n_1 n_2} x_1 x_2}{\sum_{x_1+x_2=n}^{n_1 n_2} x_1 x_2} \]

\[ = \frac{\sum_{x_1=0}^{n-1} x_1 (2-x_1)^{n-1-n_1}}{\sum_{x_1=0}^{n} x_1 (2-x_1)^{n-1}} \]

since \( x_1 + x_2 = m = 2 \),

\[ = \frac{x_1^n (2-x_1)^n}{x_1^{n+1} - (2-x_1)^{n+1}}, \] 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 FMSs. 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 efficient algorithm [1973]. 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 results, which will be used later to prove the main result.

**Proposition 1.** The set \( \chi \) of feasible loadings is a closed S-convex set.

Proof. From equation (1), we have \( \chi = \{ (x_1, \ldots, x_m) \mid \sum_{i=1}^{m} x_i = m, x_i \in \mathbb{R}, 0 \leq x_i \leq m \} \). Therefore, \( \chi \) is clearly closed. It is also clearly convex
and symmetric. Then by T11, \( \chi \) is S-convex. ||

**Proposition 2.** The quotient of two symmetric functions in the same variables
on the same symmetric set \( \chi \) is a symmetric function.

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

\[
\begin{align*}
  f(xP) &= f(x) \\
  g(xP) &= g(x)
\end{align*}
\]

for all \( x \in \chi \) and for any permutation matrix \( P \). Let

\[
h(x) = \frac{f(x)}{g(x)}
\]

for all \( x \in \chi \) such that \( g(x) \neq 0 \). Then

\[
\begin{align*}
  h(xP) &= \frac{f(xP)}{g(xP)} \\
        &= \frac{f(x)/g(x)}{g(x)} \\
        &= h(x).
\end{align*}
\]

Therefore, \( h(x) \) is a symmetric function on \( \chi \). ||

In Theorems 3 and 4 below, we prove directly the S-concavity of both the
normalizing constant, \( G(m,n;X) \), and the production function for two machines,
\( Pr(2,n;X) \).
Theorem 3. \( G(m,n;X) \) is strictly \( S \)-concave.

Proof. By \( T_\circ \) it is sufficient to show that

\[
(X_2 - X_1) \frac{\partial^2 G(m,n;X)}{\partial x_2 \partial x_1} - \frac{\partial G(m,n;X)}{\partial x_1} < 0,
\]

for \( X_1 \neq X_2 \neq 0 \).

The quantity on the left-hand side of the inequality is

\[
(X_2 - X_1) \sum_{n_1 + \ldots + n_m = n} \left( \frac{n_2}{X_2} - \frac{n_1}{X_1} \right) \frac{n_1 n_2 \ldots n_m}{X_1 X_2 \ldots X_m}
\]

\[
= \frac{X_2 - X_1}{X_1 X_2} \sum_{n_1 + \ldots + n_m = n} (n_2 x_1 - n_1 x_2) \frac{n_1 n_2 \ldots n_m}{x_1 x_2 \ldots x_m}
\]

Consider the two cases: either \( n_1 = n_2 \) or \( n_1 \neq n_2 \).

Since \( X_i > 0, i = 1, \ldots, m \), if \( n_1 = n_2 \), then the quantity is negative and we are done.

Suppose \( n_1 \neq n_2 \). Then for every \( n_1, n_2 \) such that \( n_1 > n_2 \) (say \( n_1 - n_2 = \delta > 0 \)), by symmetry there exists a corresponding pair \( n'_1, n'_2 \) such that

\[
n'_1 < n'_2 \text{ and } n'_1 - n'_2 = -\delta.
\]

In fact, \( n_1 = n'_2 \) and \( n_2 = n'_1 \). Therefore, all terms vanish except those states such that \( n_1 = n'_2 \), and we are left with:

\[
\frac{-(X_2 - X_1)^2}{X_1 X_2} \sum_{n_1 + \ldots + n_m = n} \frac{n_1 n_2 \ldots n_m}{X_1 X_2 \ldots X_m} < 0,
\]

for all \( X > 0 \) such that \( X_1 \neq X_2 \). Therefore, \( G(m,n;X) \) is \( S \)-concave.

Theorem 4. \( Pr(2,n;X) \) is \( S \)-concave.

Proof. From (4), we have:
\[
\Pr(2,n;X) = \frac{x_1^n - (2-x_1)^n}{x_1^{n+1} - (2-x_1)^{n+1}}
\]
\[
= \frac{x_1^n - x_2^n}{x_1^{n+1} - x_2^{n+2}}, \quad \text{for } X_1 \neq X_2;
\]
\[
= n/(n+\alpha-1), \quad \text{for } X_1 = X_2.
\]

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,
\[
(X_2-X_1)(\frac{\partial \Pr(2,n;X)}{\partial x_2} - \frac{\partial \Pr(2,n;X)}{\partial x_1})
\]
\[
= \frac{(X_2-X_1)([(n+1)x_2^n(x_1^n-x_2^n)-nx_2^{n-1}(x_1^{n+1}-x_2^{n+1})]-[nx_1^n(x_1^{n+1}-x_2^{n+1})-(n+1)x_1^n(x_1^n-x_2^n)])}{(x_1^{n+1}-x_2^{n+1})^{2}}
\]

Since the denominator is positive for \(X_1 \neq X_2\), it may be dropped,
yielding after simplification:
\[
(X_2-X_1)(x_2^n-1)[(n+1)x_2^n(x_1^n-x_2^n)-n(x_1^{n+1}-x_2^{n+1})]+x_1^n[(n+1)x_1^n(x_1^n-x_2^n)-n(x_1^{n+1}-x_2^{n+1})],
\]
which upon rearranging
\[
= (X_2-X_1)(X_2^n-x_2^n)(\sum_{i=1}^{n-1} x_1^{2n-2-2i}x_2^{2i}x_1^{n-1}x_2^{n-1}). \quad (5)
\]

In order to show that equation (5) is not positive, it suffices to show that
\[
x_1^{2n-2-2i}x_2^{2i}x_1^{2n-2-2i}x_2^{2i}x_1^{n-1}x_2^{n-1} \geq 2x_1^{n-1}x_2^{n-1}, \quad (6)
\]
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:

\[
x_1^{2i} x_2^{2i} (x_1^{2n-2-4i-2x_1^{n-1-2i}} x_2^{n-1-2i} + x_2^{2n-2-4i})
\]

\[
= x_1^{2i} x_2^{2i} (x_1^{2(n-1-2i)} - 2x_1^{n-1-2i} x_2^{n-1-2i})
\]

\[
= x_1^{2i} x_2^{2i} (x_1^{n-1-2i} - x_2^{n-1-2i})^2
\]

\[\geq 0.\]

Therefore, (6) holds for all \(i < n-1\). The proof for \(2i > 2n-2\) follows \textit{mutatis mutandis} and equality holds if \(i=n-1\). ||

Next consider the \(m\) machine case.

\underline{Lemma 5.} \(\Pr(m,n;X)\) is symmetric.

\textit{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.

\[
G(m,n;X) = \sum_{n_1 + \ldots + n_m = m} x_1^{n_1} x_2^{n_2} \ldots x_m^{n_m}.
\]

By Theorem 3, \(G(m,n;X)\) is strictly \(S\)-concave. Then \(G(m,n;X)\) is symmetric by T9. 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.

\underline{Theorem 6.} The production function, \(\Pr(m,n;X)\), is strictly quasiconcave.

\textit{Proof.} The result is proven in Stecke [1983b]. ||
3. CHARACTERIZING OPTIMAL WORKLOADS

We now state and prove the main result.

Theorem 7. 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, \( \chi \), is closed and S-convex.

By Lemma 5, \( Pr(m,n;X) \) is symmetric.

By \( T12 \), since \( Pr(m,n;X) \) is quasiconcave by Theorem 6, then \( Pr(m,n;X) \) is S-concave.

By \( T15 \), the set \( \chi^* \) of points maximizing \( Pr(m,n;X) \) over the set \( \chi \) is a closed S-convex set.

\( \chi^* \) is not empty since \( Pr(m,n;X) \in [0,1] \) for all \( m, n \), and \( X \in [0,m] \).

The symmetric point of \( \chi \) is the point \( [1,1,\ldots,1] \).

By \( T14 \), \( [1,\ldots,1] \subset \chi^* \).

Therefore, a balanced allocation maximizes the expected production. 

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

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 \( X \) 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 calculated and plotted, was attained at \( X = [1,1] \). Table I displays the calculated optimal allocation ratios and the maximum normalized production rate for each \( n \). For \( q_1 = q_2 = .5 \) (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
Figure 2

Production Rate as a Function of Workload Assigned to Machine 1 for 2-Machine Systems.
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.

**TABLE I**

<table>
<thead>
<tr>
<th>n</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>Max (Bal) Production Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.800</td>
</tr>
<tr>
<td>5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.833</td>
</tr>
<tr>
<td>6</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.857</td>
</tr>
<tr>
<td>7</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.875</td>
</tr>
<tr>
<td>8</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.889</td>
</tr>
<tr>
<td>9</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.900</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.909</td>
</tr>
<tr>
<td>11</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.917</td>
</tr>
<tr>
<td>12</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.923</td>
</tr>
<tr>
<td>13</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.929</td>
</tr>
<tr>
<td>14</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.933</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.000</td>
</tr>
</tbody>
</table>

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 II.
Production Rate as a Function of Workload Assigned to Machine 1 for 3-Machine Systems.
TABLE II

Maximum (Balanced) Production Rates and Corresponding Workloads for Three-Machine Systems

<table>
<thead>
<tr>
<th>n</th>
<th>Balanced Production Rate</th>
<th>Maximum Production Rate</th>
<th>$X_1^* = t_1$</th>
<th>$X_2^* = t_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>.667</td>
<td>.667</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>5</td>
<td>.714</td>
<td>.714</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>6</td>
<td>.750</td>
<td>.750</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>7</td>
<td>.778</td>
<td>.778</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>8</td>
<td>.800</td>
<td>.800</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>9</td>
<td>.818</td>
<td>.818</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>10</td>
<td>.833</td>
<td>.833</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>11</td>
<td>.846</td>
<td>.846</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>12</td>
<td>.857</td>
<td>.857</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>13</td>
<td>.867</td>
<td>.867</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>14</td>
<td>.875</td>
<td>.875</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

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 III 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

TABLE III

Maximum (Balanced) Production Rates and Corresponding Workloads for Four-Machine Systems

<table>
<thead>
<tr>
<th>n</th>
<th>Balanced Production Rate</th>
<th>Maximum Production Rate</th>
<th>$X_1^* = t_1$</th>
<th>$X_2^* = t_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>.571</td>
<td>.571</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>5</td>
<td>.625</td>
<td>.625</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>6</td>
<td>.667</td>
<td>.667</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>7</td>
<td>.700</td>
<td>.700</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>8</td>
<td>.727</td>
<td>.727</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>9</td>
<td>.750</td>
<td>.750</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>10</td>
<td>.769</td>
<td>.769</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>11</td>
<td>.786</td>
<td>.786</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>12</td>
<td>.800</td>
<td>.800</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>13</td>
<td>.812</td>
<td>.812</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>14</td>
<td>.824</td>
<td>.824</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Figure 4

Production Rate as a Function of Workload
Assigned to Machine 1 for 4-Machine Systems.
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 FMSs) considered here.

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:

\[
q_i(k) = \sum_{j=1}^{m} p_{ij}(k) q(j(k)), \quad i=1, \ldots, m
\]

\[
(k) = F \text{ or } J.
\]
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 an 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 CONCERNING 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 \( \chi \to \mathbb{R} \), where \( \chi \) is a closed subset of \( \mathbb{R}^m \). We require the following Definitions (D) and Theorems (T):

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

\[
 f(\lambda x^1 + (1-\lambda)x^2) \geq \min\{f(x^1), f(x^2)\}. 
\]

**D2** 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 \( \chi \subseteq \mathbb{R}^m \) 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 \( \chi \subseteq \mathbb{R}^m \) 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** \( \chi \) is a symmetric set if \( x \in \chi \implies x^P \in \chi \) for all permutation matrices \( P \), where \( P \) 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 \( \chi \) if for any permutation matrix \( P \)

\[
 f(x^P) = f(x) \text{ for all } x \in \chi. 
\]
D6 \( \chi \) is \( S \)-convex if \( x \in \chi \implies xS \in \chi \) for all doubly stochastic matrices \( S \), where \( S \) is a \underline{double stochastic matrix} of order \( m \) if all of its entries, \( p_{i,j} \), satisfy

i) \( p_{i,j} \geq 0 \), for all \( i, j \);

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

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

D7 \( f \) is a \underline{(strictly) \( S \)-concave function} on an \( S \)-convex set \( \chi \) if for any \( S \)

\[
f(xS) (\succ) \geq f(x) \quad \text{for all} \quad x \in \chi.
\]

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

\[
(x_2 - x_1) \left( \frac{\partial f}{\partial x_2} - \frac{\partial f}{\partial x_1} \right) (\succ) \leq 0,
\]

then the function is \underline{(strictly) \( S \)-concave} in \( \mathbb{R}^m \) (Theorem 5, p. 221, Berge [1963]).

T9 An \( S \)-concave function \( f \) in \( \mathbb{R}^m \) is \underline{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 (Theorem 3, p. 220, Berge [1963]).

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

\[
(x_2 - x_1) \left( \frac{\partial f}{\partial x_2} - \frac{\partial f}{\partial x_1} \right) (\succ) \leq 0
\]

(Theorem 6, p. 224, Berge [1963]).
Symmetric convex sets are S-convex (but not necessarily conversely).

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

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

Every nonempty S-convex set contains a symmetric point.

If \( \chi \) is a closed, S-convex set and \( f \) is S-concave on \( \chi \), then the set \( \chi^* \) of points maximizing \( f \) over \( \chi \) is a closed S-convex set (Greenberg and Pierskalla [1970]).

ACKNOWLEDGMENTS

We wish to acknowledge helpful discussions with Bruce W. Schmeiser, Leroy B. Schwarz, and James J. Solberg. The research was supported in part by National Science Foundation Grant No. APR 74 15256 to Purdue University.
REFERENCES


SOLBERG, JAMES J., "Stochastic Modeling of Large Scale Transportation Networks," Report No. DOT-ATC-79-2, School of Industrial Engineering, Purdue University, West 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, West Lafayette IN (July 1980).


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


