Division of Research
School of Business Administration
The University of Michigan

October 1990
Revised June 1991
Revised July 1992

PRODUCTION PLANNING DECISIONS IN FLEXIBLE MANUFACTURING SYSTEMS WITH RANDOM MATERIAL FLOWS

Working Paper No. 646-c

Kathryn E. Stecke
The University of Michigan
and
Narayan Raman
The University of Illinois at
Urbana-Champaign

Production Planning Decisions in Flexible Manufacturing Systems with Random Material Flows

Kathryn E. Stecke

School of Business Administration

The University of Michigan

Ann Arbor, Michigan

N. Raman

Department of Business Administration

University of Illinois at Urbana-Champaign

Champaign, Illinois

October 1990

Revised June 1991, July 1992

ABSTRACT

In this paper, we consider the FMS planning problem of determining optimal machine workload assignments in order to minimize *mean part flow time.* We decompose this problem into the subproblems of first forming machine groups and next assigning operations to these groups. Three types of grouping configurations — no grouping, partial grouping and total grouping, are considered. In both *no grouping* and *partial grouping,* each machine is tooled differently. While each operation is assigned to only one machine in no grouping, partial grouping permits multiple operation assignments. On the other hand, *total grouping* partitions the machines into groups of identically-tooled machines; each machine within a group is capable of performing the same set of operations. Within this grouping framework, we consider three machine loading objectives - minimizing the total deviation from the optimal group utilization levels, minimizing part travel and maximizing routing flexibility, for generating a variety of system configurations.

A queueing network model of an FMS is used to determine the optimal configurations and machine workload assignments for the no grouping and total grouping cases. It is shown that under total grouping, the configuration of M machines into G groups that minimizes flow time is one in which the sizes of the machine groups are maximally unbalanced and the workload per machine in the larger groups is higher. This extends previous results on the optimality of unbalancing both machine group sizes and machine workload to the mean flow time criterion.

A simulation experiment is next conducted to evaluate the alternative machine configurations to understand how their performance depends upon the system characteristics, such as utilization level and variation among operation processing times. We also study the robustness of these configurations against disruptions, such as machine unreliability and variation in processing batch sizes. While different configurations minimize mean flow time under different parameter values, partial grouping with state-dependent part routing performs well across a wide range of these values. Experimental results also show that the impact of disruptions can be reduced by several means, such as aggregating operations of a part to be performed at the same machine and maximizing the number of operation assignments (in order to minimize part movements), in addition to providing routing flexibility.

1 Introduction

Greater product proliferation and market fragmentation, and shorter product life cycles have made firms increasingly aware of the importance of manufacturing flexibility. Unlike conventional manufacturing methods, programmable automation with computer-controlled and versatile machining and assembly capabilities promises an effective solution to the simultaneous requirement of manufacturing efficiency and process flexibility. Consequently, the design and operation of flexible manufacturing systems (FMSs) and the definition and classification of production flexibility are subjects of growing interest among researchers and practitioners alike.

The manufacturing issues faced in an FMS can be categorized into: i) design problems, ii) planning problems, and iii) scheduling and control problems. FMS *design* problems address the long term issues relating to the system, and they include decisions regarding the selection of part types that can be produced in the system, selection and layout of machine tools and the material handling system, design of buffers and the computer control architecture. FMS *planning* problems comprise resource allocation decisions during pre-production system setup. They include selecting the subset of part types for imminent manufacture from the set of all part types that the FMS can produce, determining the ratios in which these part types will be manufactured concurrently, the assignment of pallets, fixtures, etc. to these part types and the allocation of operations and tools to individual machines. FMS *scheduling and control* problems relate to the execution of orders and include the determination of part input sequences, the part processing sequence at each machine and monitoring the actual system performance and taking the necessary corrective actions.

Within such a hierarchical categorization of FMS problems, this study addresses the planning level. Much of the effectiveness of an FMS is derived from the fact that a part can have, in general, several alternative routes through the system. The ability of an operation to select a machine in real time based on the current system status reduces part flow time relative to a conventional system in which each operation is typically assigned to only one machine. Routing flexibility also renders the system less susceptible to disruptions such as machine failures. The number of such alternative part routes is determined by the manner in which the operations of individual part types are assigned to the various machines — a decision that is made at the planning level, and the resulting machine grouping configuration of the system.

The performance measure of interest in this study is *mean part flow time.* In an FMS, mean flow time includes machining time (processing time and waiting time at machines) as well as material transfer time (travel time plus waiting time for the transporters). This work focuses on the machining times. The objectives of this paper are to generate and evaluate alternative machine configurations of a dynamic FMS as well as to understand how their relative performance depends upon many underlying system characteristics. These configurations are generated by solving the *machine grouping* and *machine loading* problems.

Following Stecke (1983) and Stecke and Solberg (1985), the machine grouping problem is the problem of optimally partitioning the available machines into groups of identically-tooled machines and determining the appropriate machine workloads. We investigate three types of grouping configurations – no grouping, partial grouping and total grouping. The terminology used here to define various FMS configurations is as follows. In both *no grouping* and *partial grouping*, each machine is tooled differently. While each operation is assigned to only one machine in no grouping, partial grouping permits multiple operation assignments. On the other hand, *total grouping* partitions the machines into groups of identically-tooled machines. Each machine within a group is capable of performing the same set of operations and different groups have mutually disjoint operation processing capabilities. If *all* machines are identically-tooled, this set of machines is called a *pool.*

In the first stage, we develop an open queueing network model for no grouping and total grouping in order to determine characteristics of optimal solutions. These characteristics are then incorporated within a mathematical programming model for the machine loading problem. The objective of the machine loading problem is to assign operations to machines such that the resulting machine workloads conform closely to the optimal workloads determined by the solution to the grouping problem. In addition, this study considers two secondary loading objectives – minimizing part travel and maximizing operation routing flexibility. These loading objectives are combined with the three grouping types discussed above to generate several system configurations.

At the next stage, a simulation study is used to evaluate these configurations. The purpose of these experiments is three-fold. A first is to better understand the case of partial grouping because these configurations are not easily amenable to analytical evaluation using queueing networks. Second, we study the impact of relaxing some of the assumptions made while developing the queueing network model. Third, we evaluate the robustness of each configuration on the basis of two criteria — the insensitivity of the configuration to the actual scheduling rule used and the deterioration of system performance in the face of disruptions. Conway et al. (1967) observe that one of the major benefits of providing routing flexibility is that the system is less sensitive to schedule quality. Consequently, one of the major objectives at the FMS planning stage is to simplify the decisions that need to be made at the subsequent scheduling and control stage. In evaluating the second criterion of robustness, we consider two types of disruptions — machine breakdowns and variations in the batch size of a given part type.

Much of the previous research on dynamic FMSs is based on queueing-theoretic approaches. Buzacott and Yao (1986) present an excellent survey of this literature. Studies that address closely related issues include Stecke (1983, 1986a, 1986b), Stecke and Solberg (1985), Shanthikumar and Yao (1987, 1988), Stecke and Kim (1989, 1991), Dallery and Stecke (1990), and Arbib et al. (1991). Some of these investigations model an FMS as a closed network of multiserver queues and derive the optimal partitioning of the system into groups of identically-tooled machines for the objective of maximizing system throughput. The underlying result of Stecke and Solberg (1985) is that under certain service disciplines and operation processing time distributions, the expected part production rate is maximized by grouping machines into unequally sized groups and assigning appropriately unbalanced workloads per machine to these groups. Shanthikumar and Yao (1987, 1988) show that the throughput of a machine group is concave in the number of machines. This results in an efficient heuristic algorithm for assigning a given number of machines to individual groups. Stecke (1986a) shows that throughput is quasiconcave in the workload allocated per machine for single-machine groups. Stecke (1983, 1986b) considers various operation assignment objectives appropriate in an FMS and presents a hierarchical framework for considering these objectives.

There is a parallel body of research which addresses the allocation of operations to machines in a *static* FMS environment. Ammons, Lofgren and McGinnis (1984), Kusiak (1984), Rajagopalan (1986), Berrada and Stecke (1986), and Hwang (1986) present mathematical programming approaches to solve this problem for various objectives. However, because of the static nature of the problem considered, they do not explicitly address the impact of system utilization levels and unexpected disruptions.

This work differs from the previous studies on dynamic FMSs in the following aspects. First, we extend the notion of the optimality of unbalancing machine workloads to the performance measure of mean flow time. To our knowledge, this has not been done before. These results are then used to link the solutions to the machine grouping and machine loading problems in a hierarchical manner. Second, we address the issue of robustness; in particular, we study the impact of schedule quality on system configurations. Finally, we demonstrate the effectiveness of *partial grouping* under ideal conditions as well as its robustness against disruptions. Partial grouping configurations have not been previously investigated to our knowledge. Results of this paper indicate that these configurations can yield significant performance improvements. This study also helps to clarify the different circumstances in which each of the various types of configurations investigated is best for minimizing mean part flow time.

The paper is organized as follows. In §2, we develop a general formulation of the minimum mean flow time problem. In view of the complexity of this problem, it is decomposed heuristically into the machine grouping and the machine loading problems. In §3, we model an FMS as an open network of $M/M/c$ queues to address the machine grouping problem and derive the characteristics of an optimal solution. The machine loading problem is discussed in §4. The experimental investigation is presented in §5. We conclude in §6 with a summary of the main results obtained in this paper.

2 FMS Planning Problem

Consider an FMS consisting of M machines. Let N be the number of different part types produced in the system. We assume that all machines are of the same type because machines are grouped only within a particular machine type. Orders for these part types arrive randomly to the system. A particular part type requires a series of operations to be performed in a specified sequence. Each operation can be assigned to one or more machines capable of processing that operation by ensuring that the required tools are available at the machines. Each machine, say m, has a tool magazine of limited tool slot capacity T^m. We note that there are FMSs with tool delivery capabilities, in which cutting tools can be interchanged between the machine and a central tool storage facility automatically in real time. At the planning level, this can result in a virtually unlimited tool magazine capacity (unless too many tools require transport or automatic change at the same time). Operation assignment in such systems is a real-time decision that is done simultaneously with operation and transporter scheduling. We do not consider these types of systems in this study.

One objective of the production planning problem is to assign operations to machines such that the steady-state mean part flow time is minimized. This problem is formulated below; the notation used here is given in Table 1.

INSERT TABLE 1 HERE

MFT1

$$\min \frac{1}{\lambda} \sum_{m=1}^{M} \sum_{j=1}^{N} \sum_{i=1}^{n_j} \lambda^j W_{ijm} x_{ijm} \tag{1}$$

subject to

$$\sum_{j=1}^{N} \sum_{i=1}^{n_j} t_{ij} x_{ijm} \leq T^m, \ \forall m \tag{2}$$

$$\rho_m = \sum_{j=1}^{N} \sum_{i=1}^{n_j} p_{ij} \lambda^{ijm} x_{ijm}, \ \forall m \tag{3}$$

$$\sum_{m=1}^{M} \lambda^{ijm} = \lambda^j, \ \forall i,j \tag{4}$$

$$\lambda^{ijm} = f(x_{ijm}, \rho_m, \lambda^j), \ \forall i,j,m \tag{5}$$

$$0 \leq \rho_m < 1, \ \forall m \tag{6}$$

$$x_{ijm} \in \{0,1\}, \ W_{ijm}, \lambda_{ijm} \geq 0 \ \ \forall i,j,m \tag{7}$$

Equation (1) expresses the expected part flow time as the weighted sum of the average time spent by each operation at the machine(s) to which it is assigned; W_{ijm} is the sum of the expected waiting and processing times for operation i of job j at a machine m. Equation (2) relates to the constraints on the tool magazine capacity. Equation (3) measures the machine utilizations that result from a given assignment of operations to machines. Equation (4) ensures that the sum of operation arrival rates λ^{ijm} at different machines equals the part arrival rate. Equation (5) expresses the dependence of these operation arrival rates upon the assignment variables x_{ijm}, machine utilizations ρ_m, and the part arrival rates λ^j. Finally, Equations (6) and (7) specify the range of valid machine utilizations and the nature of the problem variables, respectively.

We assume here that the cutting tools are not shared among different operation types. This assumption is made primarily to keep the model relatively simple. Relaxing it will introduce additional complexity without adding to our understanding of the particular issues under study here, for example, the underlying differences among the various system configurations. In addition, all machines are considered to be equally efficient in terms of the operation processing times. In general, machines of the same type are equally efficient.

MFT1 is a hard nonlinear 0-1 programming problem. The major difficulty arises in characterizing W_{ijm} in a network of $G/GI/c$ queues. A related issue is that of specifying the functional form of Equation (5). In order to solve this problem, we decompose it heuristically into the machine grouping and the machine loading problems, which are solved sequentially.

3 Machine Grouping Problem

In the machine grouping problem **MGP**, we determine the optimal partition of machines into groups. This involves: 1) determining the number of groups G, and 2) assigning machines to individual groups. In §3.1 we discuss total grouping using queueing networks. These results are extended to the case of no grouping in §3.2. Partial grouping is discussed in §3.3. These are all evaluated and compared in §4.

3.1 Total Grouping

The minimum mean flow time problem is restated for the total grouping configuration as **MFT2** by taking groups into account.

MFT2

$$\min \frac{1}{\lambda} \sum_{g=1}^{M} \sum_{j=1}^{N} \sum_{i=1}^{n_j} \lambda^j W_g x_{ijg} \tag{8}$$

subject to

$$\sum_{g=1}^{M} x_{ijg} = 1, \ \forall i, j \tag{9}$$

$$\sum_{g=1}^{M} y_{mg} = 1, \ \forall m \tag{10}$$

$$\sum_{j=1}^{N} \sum_{i=1}^{n_j} t_{ij} x_{ijg} \leq T^m y_{mg} + B(1 - y_{mg}), \ \forall m, g \tag{11}$$

$$\sum_{j=1}^{N} \sum_{i=1}^{n_j} t_{ij} x_{ijg} \leq B(\sum_{m=1}^{M} y_{mg}), \ \forall g \tag{12}$$

$$\rho_g = \frac{\sum_{j=1}^{N} \lambda^j \sum_{i=1}^{n_j} p_{ij} x_{ijg}}{\sum_{m=1}^{M} y_{mg}} \ \forall m, g \tag{13}$$

$$0 \leq \rho_g < 1, \ \forall g \tag{14}$$

$$x_{ijg}, y_{mg} \in \{0, 1\}, \ \forall i, j, m, g \tag{15}$$

This formulation introduces the additional variable y_{mg}, which equals 1 if machine m is assigned to group g, and is zero otherwise. Equation (9) ensures that each operation is assigned to only one group; it replaces Equation (4) in **MFT1**. Equation (10) ensures that each machine is assigned to only one group. Equation (11) specifies that an operation is assigned to a group only if every machine in that group has adequate tool magazine capacity available. Equation (12) ensures that no operation is assigned to a group to which no machines are assigned. Equation (13) determines the utilization of each machine within a group. Equations (14) and (15) parallel Equations (8) and (9) in **MFT1** with subscript m replaced by g wherever appropriate.

MGP uses **MFT2** at an aggregate level by combining all part types into a single aggregate part type. At this aggregate level, an open queueing network is used to model the

FMS. Let the FMS comprise G groups such that group g consists of m_g machines. Clearly $G \leq M$ and $\sum_{g=1}^{G} m_g = M$. Let the transition probability that a part that has completed processing at group i will next visit a machine in group g be given by π_{ig}. The probability that the part will exit the system after finishing at group i is $1 - \sum_{g=1}^{G} \pi_{ig}$. Let λ be the external part arrival rate to the system and λ_g be the average part arrival rate to group g. From traffic balance, we have

$$\lambda_g = \gamma_g + \sum_{i=1}^{G} \lambda_i \pi_{ig}, \;\; g = 1, \ldots, G$$

where γ_g is the external arrival rate to group g.

We define α_g, the visit ratio at group g, as the average number of times that the aggregate part type is processed at group g. Only one operation at a time is processed at each visit to a group. In addition, let $1/\mu_g$ be the average processing time of a part at group g, $1/\mu$ the average part processing time, L_g the average number of parts at group g, and T_g the maximum number of operations that can be assigned to group g.

We have the following identities:

$$\rho_g = \frac{\lambda_g}{m_g \mu_g},$$

$$\alpha_g = \lambda_g / \lambda = m_g \rho_g \mu_g / \lambda$$

$$\frac{1}{\mu} = \sum_{g=1}^{G} \alpha_g / \mu_g = \frac{1}{\lambda} \sum_{g=1}^{G} m_g \rho_g. \tag{16}$$

We also have the following relationships between the variables in **MFT2** and **MGP**:

$$\alpha_g = \frac{\sum_{j=1}^{N} \lambda^j \sum_{i=1}^{n_j} x_{ijg}}{\lambda}, \;\; \forall g$$

$$m_g = \sum_{m=1}^{M} y_{mg}, \;\; \forall g$$

$$\frac{1}{\mu_g} = \frac{\sum_{j=1}^{N} \lambda^j \sum_{i=1}^{n_j} p_{ij} x_{ijg}}{\sum_{j=1}^{N} \lambda^j \sum_{i=1}^{n_j} x_{ijg}}, \;\; \forall g.$$

In addition, because all machines within a group are tooled identically, the maximum number of operations T_g that can be assigned to group g is given by

$$T_g = \min_{m} \{ T^m \mid y_{mg} = 1 \}.$$

From Little's law, the expected part flow time can now be restated as

$$\begin{aligned} MFT &= \frac{1}{\lambda}\sum_{g=1}^{G} W_g \sum_{j=1}^{N} \lambda^j \sum_{i=1}^{n_j} x_{ijg} \\ &= \sum_{g=1}^{G} W_g \alpha_g \\ &= \frac{1}{\lambda}\sum_{g=1}^{G} L_g. \end{aligned}$$

For a given G, the machine grouping problem **MGP**$_G$ can then be formulated as:

MGP$_G$

$$\min \quad MFT = \frac{1}{\lambda}\sum_{g=1}^{G} L_g(m_g, \rho_g) \tag{17}$$

subject to

$$\sum_{g=1}^{G} \alpha_g/\mu_g = 1/\mu \tag{18}$$

$$t_{ave}\alpha_g \leq T_g, \ \forall g \tag{19}$$

$$\alpha_g = m_g \rho_g \mu_g / \lambda \tag{20}$$

$$\sum_{g=1}^{G} m_g = M \tag{21}$$

$$0 \leq \rho_g < 1, \ \ \alpha_g \geq 0, \ \ L_g \geq 0, \ \ m_g \geq 0, \text{integer} \ \ \forall g. \tag{22}$$

where t_{ave} is the average number of tool slots required by an operation. Using Equation (16), we can rewrite Equation (18) as

$$\sum_{g=1}^{G} m_g \rho_g = \lambda/\mu = M\rho, \tag{23}$$

where $\rho = \lambda/(M\mu)$ is the overall system utilization. Note that the right hand side of Equation (23) is a constant for planning level decisions.

3.1.1 Total Grouping Results

Even for known values of G, MGP remains quite difficult to solve primarily because of the cumbersome expression relating L_g to ρ_g and m_g. However, the following propositions indicate that a sequential solution approach can be constructed efficiently.

Proposition 1. *MFT does not decrease if any group g is decomposed into two subgroups $g1$ and $g2$, while the other groups remain unchanged.*

PROOF. See Appendix 1.

Proposition 1 directly leads to the following result.

Proposition 2. *MFT is minimized by minimizing the total number of groups.*

This result holds independently of how the machines are partitioned into groups, and how the workloads are allocated. Consequently, our solution approach to **MGP** consists of first finding the optimal number of groups G^*. At the next step, we solve Problem $\mathbf{MGP}_{G^*}$ to obtain the optimal partition $\mathbf{m}^* = (m_1^*, m_2^*, \cdots, m_{G^*}^*)$ of machines into groups and in so doing, we also determine the optimal utilizations of individual machines in each group, $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{G^*}^*)$.

In order to develop some characteristics of the optimal solution, we determine the optimal utilization levels ρ^* corresponding to specific (feasible) values of $\mathbf{m}$. First, consider the case in which $m_1 = m_2 = \cdots = m_G$, for which we have the following result.

Proposition 3. *MFT is minimized for a system of machine groups of equal sizes by allocating balanced workloads to all machines in each group.*

PROOF. Note that L_g is convex in ρ_g (Lee and Cohen 1983) and therefore, for equal-sized groups, MFT is a sum of identical convex functions. □

However, for unequal-sized groups, the optimal group utilizations will depend on the number of machines in each group. In the following, we determine these utilizations for 3-, 4- and 5-machine systems by initially ignoring tool magazine capacity restrictions and extend these results to the general system through Conjecture 1.

Three machines can be grouped in two ways: (1, 1, 1) and (1, 2). From Proposition 3 it follows that MFT is minimized in the (1, 1, 1) configuration by providing equal machine utilizations, which are given by

$$\rho_g^* = \rho = \frac{\lambda}{3\mu}, \quad g = 1, 2, 3$$

and the resulting minimum mean flow time is

$$MFT^*(1,1,1) = \frac{1}{\lambda}\sum_{g=1}^{3} L_g = \frac{1}{\lambda}\sum_{g=1}^{3} \frac{\rho_g^*}{1-\rho_g^*} = \frac{3\rho}{\lambda(1-\rho)}.$$

For the (1, 2) configuration, the minimum MFT is given by

$$MFT^*(1,2) = \frac{1}{\lambda}\left[\frac{\rho_1^*}{1-\rho_1^*} + \frac{2\rho_2^*}{1-\rho_2^{*2}}\right].$$

In Appendix 2 it is shown that ρ_1^* and ρ_2^* are obtained by solving

$$(4-12\rho)\rho_2^* + (7-6\rho+9\rho^2)(\rho_2^*)^2 + (4-12\rho)(\rho_2^*)^3 + 3(\rho_2^*)^4 - 6\rho + 9\rho^2 = 0,$$

and

$$\rho_1^* = 3\rho - 2\rho_2^*.$$

The resulting machine utilizations, ρ_1^* and ρ_2^*, are shown in Figure 1 for various values of system utilization ρ. Note that $\rho_2^* > \rho_1^*$ for all ρ. In addition, as $\rho \rightarrow 1$, ρ_1^* and $\rho_2^* \rightarrow \rho$. Hence, MFT is minimized by utilizing each machine in the larger group more heavily. However, the imbalance between machine utilizations of the two groups decreases with an increase in the overall system utilization. In the limit, both groups have the same utilization.

INSERT FIGURE 1 HERE

Figure 2 compares the MFT values under the optimal (1, 1, 1) and (1, 2) configurations for various values of ρ. Note that the (1, 2) configuration with appropriately unbalanced workloads consistently results in smaller MFT. Furthermore, the difference between these two configurations grows at an increasing rate as the overall system utilization increases.

INSERT FIGURE 2 HERE

Four machines can be grouped in 4 ways — (1, 1, 1, 1), (1, 1, 2), (2, 2) and (1, 3). Similarly, the alternative configurations possible in a 5-machine system are (1, 1, 1, 1, 1), (1, 1, 1, 2), (1, 2, 2), (1, 1, 3), (2, 3) and (1, 4). Figures 3 and 4 depict the MFT values obtained under these configurations given optimal group utilization levels for 4- and 5-machine systems, respectively. These figures extend the result obtained earlier for the 3-machine system. In addition, they bring out the relative impact of fewer groups and unequal group utilizations individually.

INSERT FIGURES 3 AND 4 HERE

Consider, for example, the 5-machine system. MFT decreases as the number of groups decreases from 5 to 2. For a given number of groups, MFT is minimized by maximally unbalancing the group sizes. For instance, (1, 1, 3) is superior to (1, 2, 2) when $G = 3$. Similarly, $(1,4)$ is superior to $(2,3)$ when $G = 2$. Figures 3 and 4 also show that reducing the number of groups is more effective than unbalancing group sizes and allocating appropriately unbalanced workloads.

These results lead to the following conjecture.

Conjecture 1. *MFT is minimized by minimizing the number of machine groups, grouping machines into unequal groups, and by allocating appropriately unbalanced workloads to these groups.*

Because of the cumbersome nature of the MFT function, it is difficult to verify the generality of this assertion. However, it has been proved to be true for the several systems that we have examined. Proposition 2 and Conjecture 1 parallel the conjectures stated in Stecke and Solberg (1985), who studied the production rate function in closed queueing networks of multiserver queues.

3.1.2 Machine Grouping

The solution method for **MGP** follows from Conjecture 1: the individual steps in the procedure are: 1) determining the minimum number of groups that can be formed, 2) allocating the available machines to these groups such that these group sizes are maximally unbalanced, and 3) determining the appropriate groups utilization levels. These steps are discussed below.

If all machines have the same tool magazine capacity T, then the minimum number of machine groups required is given by

$$G^* = \left\lceil \frac{\sum_{j=1}^{N} \sum_{i=1}^{n_j} t_{ij}}{T} \right\rceil,$$

where $\lceil a \rceil$ is the smallest integer greater than or equal to a.

If individual machines have different tool magazine capacities, G^* can be found by applying the following procedure. Renumber all machines in the non-increasing order of T^m. Then, G^* is the smallest integer K such that

$$\frac{\sum_{j=1}^{N} \sum_{i=1}^{n_j} t_{ij}}{\sum_{l=1}^{K} T^l} \leq 1.$$

Note that this step insures that the groups formed are feasible with respect to Equation (19). The optimal grouping configuration is given by

$$\mathrm{m}^* = (1, 1, \cdots, M - G^* + 1).$$

The optimal group utilizations are obtained by solving the following problem.

$$\min \ \frac{1}{\lambda}\Big[\sum_{g=1}^{G^*-1} \frac{\rho_g}{1-\rho_g} + L(M - G^* + 1, \rho_{G^*})\Big]$$

subject to

$$(G^* - 1)\rho_g + (M - G^* + 1)\rho_{G^*} = \lambda/\mu$$

$$0 \leq \rho_g < 1, \ \forall g.$$

Because G^* is known, $L(M - G^* + 1, \rho_{G^*})$ can be expressed as a function of only ρ_{G^*}. This problem can be solved in a manner similar to the 3-, 4- and 5-machine systems discussed earlier. The resulting ρ^* is next input into the Machine Loading Problem **MLP**.

3.2 No Grouping

Because no two machines are tooled alike in the case of no grouping, the system comprises M single machine groups. From Proposition 3, it follows that machine workloads are balanced in the optimal configuration. Consequently, $\rho_m = \rho, \ \forall m$.

3.3 Partial Grouping

Since each machine is tooled differently, M single machine groups are formed in partial grouping as well; consequently, the results of Proposition 3 apply. However, in this case, the fact that an operation can be performed at more than one machine permits state-dependent routing (Towsley 1980, Sauer 1983, Yao and Buzacott 1985) which could perform better than the case in which the selection of part routes is based on pre-determined branching probabilities (Chow and Kohler 1979). Analytical performance evaluation of such a policy for the FMS type considered here using queueing networks is, however, difficult. One of the major difficulties lies in efficiently decomposing the FMS into mutually exclusive subnetworks consisting of parallel machines. It is unclear if there is any particular merit to unbalancing machine workloads in such configurations. On the other hand, studies of single-stage systems by Winston (1977) and Chow and Kohler (1979) indicate that a policy which routes a part to the machine with the shortest queue is likely to perform well. Such a policy helps to level machine workloads.

4 Machine Loading Problem

Given the solution to **MGP**, the machine loading problem deals with the allocation of operations to individual groups such that deviations of actual utilizations from their ideal values are minimized. This leads to the following formulation for the case of total grouping.

MLP

$$\min \sum_{g=1}^{G^*} \mid \rho_g - \rho_g^* \mid$$

subject to

$$\sum_{g=1}^{G^*} x_{ijg} = 1, \;\; \forall i,j \tag{24}$$

$$\sum_{j=1}^{N} \sum_{i=1}^{n_j} t_{ij} x_{ijg} \leq T_g, \;\; \forall g \tag{25}$$

$$\rho_g = \frac{\sum_{j=1}^{N} \lambda^j \sum_{i=1}^{n_j} p_{ij} x_{ijg}}{m_g}, \;\; \forall g \tag{26}$$

$$0 \leq \rho_g < 1, \;\; \forall g \tag{27}$$

$$x_{ijg} \in \{0,1\}, \;\; \forall i,j,g \tag{28}$$

For the cases of no grouping and partial grouping, we replace subscript g with m. In addition, for partial grouping we substitute Equation (24) with

$$L_{ij} \leq \sum_{m=1}^{G^*} x_{ijm} \leq U_{ij}, \;\; \forall i,j \tag{29}$$

to account for any pre-specified lower bound L_{ij} and upper bound U_{ij} on the number of permissible assignments for any operation. [If these bounds are not specified, then trivially $L_{ij} = 1$, and $U_{ij} = B$, where B is a large number.]

A polynomial time exact algorithm for solving this problem is unlikely to exist because it can be shown to be NP-complete. We propose a heuristic solution approach which is a modification of the first fit decreasing heuristic for the bin packing problem. The algorithm consists of the following steps:

1. a) Determine the target workloads $\theta_g,\ g = 1, 2, \ldots, G^*$.

$$\theta_g = m_g \rho_g^*,\ g = 1, 2, \ldots, G^*.$$

b) Initialize the counters for the current workload W_g, the remaining assignable workload Δ_g, and the remaining tool magazine capacity τ_g, for each group.

$$\begin{aligned} W_g &= 0,\ g = 1, 2, \ldots, G^* \\ \Delta_g &= \theta_g,\ g = 1, 2, \ldots, G^* \\ \tau_g &= \min_{m \in g}\{T^m\},\ g = 1, 2, \ldots, G^*. \end{aligned}$$

c) Form two lists of operations. For no grouping and total grouping, the primary list consists of one copy of each operation, and the secondary list is empty. For partial grouping, the primary list consists of L_{ij} copies, and the secondary list consists of $U_{ij} - L_{ij}$ copies of each operation. Arrange all operations in both lists in the decreasing order of $w_{ij} = \lambda_j p_{ij}$.

2. Assign the operation i^*j^* at the head of the list of unassigned operations in the primary list to g^*, where

$$g^* = arg\,max_g\{\frac{\Delta_g}{m_g} \mid \tau_g \geq t_{i^*j^*}\}.$$

Update workloads and available tool magazine capacity.

$$\begin{aligned} W_{g^*} &\leftarrow W_{g^*} + w_{i^*j^*}, \\ \Delta_{g^*} &\leftarrow \Delta_{g^*} - w_{i^*j^*}, \\ \tau_g &\leftarrow \tau_g - t_{i^*j^*}. \end{aligned}$$

If $\tau_g = 0$, eliminate group g from further consideration.

3. Repeat Steps 2 and 3 until all operations in the primary list are assigned.

4. Stop in the cases of no grouping and total grouping. For partial grouping, go to Step 5.

5. With respect to operation i^*j^* at the head of the unassigned operations in the secondary list, find group g^* such that

$$\begin{aligned} x_{i^*j^*g^*} &= 0, \\ \frac{\Delta_{g^*} + w_{i^*j^*}}{m_{g^*}} &< 1, \\ \left| \frac{\Delta_{g^*} + w_{i^*j^*}}{m_{g^*}} - \rho_g^* \right| &< \left| \frac{\Delta_{g^*}}{m_{g^*}} - \rho_g^* \right|, \text{ and} \\ g^* &= arg\, max_g \{ \frac{\Delta_g}{m_g} \}. \end{aligned}$$

   If these conditions are satisfied for any group, assign operation i^*j^* to group g^* and update the workload and remaining tool magazine capacities as shown in Step 2. Otherwise, delete all copies of operation i^*j^* from the secondary list.

6. Repeat Step 5 if the secondary list is not empty. Otherwise, stop.

In many systems, it may be appropriate to consider other loading objectives in addition to the objective of ensuring appropriate group utilizations. Following Stecke (1983), we consider two such objectives. The first is minimizing part movements. This objective is particularly useful for the case in which travel times are significant and/or the material transporters are heavily utilized. In addition, this objective leads to an aggregation of operations of a given part type at a machine. Consequently, each part tends to join fewer machine queues.

The formulation of the machine loading problem corresponding to the objective of minimizing part travel, **MLPMT**, is given below.

MLPMT

$$\min \sum_{g=1}^{G^*} \sum_{j=1}^{N} \sum_{i=1}^{n_j - 1} \mid x_{ijg} - x_{i+1,jg} \mid$$

subject to

$$\mid \rho_g - \rho_g^* \mid \leq \epsilon, \ \forall g \tag{30}$$

and (24)-(28)

In this formulation, ϵ denotes the maximum deviation from the ideal value permitted to the actual utilization of any group.

The second objective considered is the maximization of the weighted number of operation assignments. The weights assigned to individual operations reflect their criticality. This objective attempts to increase system flexibility selectively by providing more alternative routes to operations that have greater impact on the overall system performance.

Determining the criticality c_{ij} of a given operation i in part type j is, however, difficult. If processing times are a measure of criticality, then relative to **MLP**, longer operations will be assigned more often under this objective. On the other hand, if all operations are considered equally critical, then this objective will lead to more duplications of the shorter operations. We consider the relative merits of these two extreme scenarios in greater detail in §5.

The loading problem corresponding to the secondary objective of maximizing flexibility, **MLPMF**, is formulated below.

MLPMF

$$\max \sum_{g=1}^{G^*} c_{ij} x_{ijg}$$

subject to

(24)-(28), (30)

We combine the three loading objectives given in formulations **MLP**, **MLPMT** and **MLPMF**, with the grouping constructs of no grouping, partial grouping and total grouping to generate system several configurations. The experimental investigation of these configurations is now discussed.

5 Experimental Study

In this section, simulation experiments are performed in order to extend our investigation to a general FMS that is based upon the facility at a major manufacturing plant in Illinois producing heavy engineering equipment. One objective of these experiments is to evaluate the relative performance of partial grouping configurations. We also test the robustness of the results obtained in the previous sections when the assumption regarding exponentially distributed operation processing times is relaxed. Specifically, parts now have deterministic processing times. In addition, we measure the effectiveness of the various grouping and loading objectives under different values of the system parameters and in the face of system disruptions.

The two system parameters that are studied are system utilization level ρ, and the coefficient of variation of the operation processing times (CVOPT). While the impact of ρ on MFT is well known, different system configurations are likely to respond differently to a change in ρ. Recent studies (see, for example, Kochman 1989, Monahan and Smunt 1990) show that mean part flow time is affected significantly by the variability in operation processing times as well. As Monahan and Smunt indicate, CVOPT can be considered as a surrogate for system disruptions. However, it merits independent consideration because it affects the coefficient of variation of service times at individual machines. An increase in CVOPT will likely result in larger MFT for any system. An important measure of the effectiveness of any configuration is its robustness against changes in CVOPT.

In addition to varying CVOPT, we consider two kinds of system disruptions. The first is machine breakdowns. The degree of unreliability of a machine can be expressed in terms of its mean time to failure (MTTF). The smaller the MTTF, the greater the unreliability. The second type of disruption that we consider is the variation in the batch size of a given part type. It is well-known (see, for example, Kleinrock 1975) that bulk arrivals, especially in varying batch sizes, result in larger MFT. Consequently, in a manufacturing system, batch size variations have a disruptive impact. Such variations are caused, for example, by fluctuations in customer order quantity. In many multi-stage manufacturing systems, process batch sizes variability is a result of changes in machine yield and/or transfer batch sizes.

5.1 Experimental Design

The experiments consider a dynamic FMS that produces twelve part types to order. Orders for these part types arrive randomly to the system following a Poisson process. Each part type requires six operations with deterministic processing times. The FMS consists of six machines. The processing time of a particular operation type is the same across all machines. Material handling, and the scheduling and routing of the material transporters are not considered here so as to not confound these issues with the machine configuration and operation allocation issues that are the focus of this study. Consequently, we assume that material transfer times are negligibly small.

Four system utilization levels — ρ = 0.6, 0.7, 0.8 and 0.9 are considered. Operation processing times are sampled from a uniform distribution to yield three levels — 0.0, 0.4, and 0.8 of CVOPT. Two scheduling rules — First-come-first-served (FCFS) and Shortest Processing Time (SPT) are used to evaluate the impact of the quality of the scheduling rule on different configurations. FCFS is used primarily to serve as a benchmark. SPT is widely regarded as an effective heuristic for the mean flow time problem. Thus, a large difference between the FCFS and SPT values for a given configuration would imply that it is very sensitive to the quality of the scheduling rule.

Three levels of unreliability are considered corresponding to MTTF values of ∞, $10\overline{p}$ and $5\overline{p}$, where $\overline{p}$ is the average part processing time. An exponential distribution is used to represent the time to the next failure for any machine. In each case, the mean time to repair a machine is sampled from a uniform distribution with mean $0.3\overline{p}$. Three levels of batch sizes are used in the study. In the first level, the batch size is fixed at 1. In the second level, the batch size is sampled from the uniform distribution (3,7), while in the third level, the batch size is sampled from the uniform distribution (6,14). Note that in the two latter cases, the ratio of the range to the mean is the same. We study the effect of machine failures and variation in batch sizes at the system utilization level of 0.9. A high utilization level is selected primarily to highlight the impact of such disruptions.

The method of replications is used to obtain the summary statistics. Each scenario is replicated four times. Within each replication, steady-state statistics are obtained for over 4500 parts.

5.2 System Configurations

While each machine has the *capability* of processing any one of the required operations, its ability to *execute* an operation in real time depends whether it has the required cutting tools. We consider the general case in which, because of the tool magazine capacity constraints, it is not possible to assign all operations of all part types to a single machine. Therefore, all machines cannot be pooled into a single group. The actual assignment of operations to machines is given by the system configuration. The various configurations studied are described below.

5.2.1 Grouping Configurations

All three grouping configurations discussed earlier, no grouping, partial grouping and total grouping, are examined. For comparison purposes, we consider three groups under total grouping. This leads to three possible configurations — (1, 1, 4) which is maximally unbalanced, (2, 2, 2) which is perfectly balanced, and the intermediate configuration (1, 2, 3).

Corresponding to a given system utilization ρ, the optimal workload per machine in each group for each configuration was determined by using the approach given in §3.1.2; the resulting values are shown in Table 2. In this table, ρ_1, ρ_2, and ρ_3 refer to the optimal utilization of each machine in groups 1, 2 and 3, respectively. For example, corresponding to $\rho = 0.2$ under the (1, 2, 3) configuration, the utilization levels that minimize mean flow time are as follows: the machine in group 1 is assigned a utilization level of 0.044, each of the two machines in group 2 is assigned a utilization of 0.173, and each of the three machines in group 3 has a utilization of 0.270. Then $\rho_1^* + 2\rho_2^* + 3\rho_3^* = 1.2 = M\rho$.

INSERT TABLE 2 HERE

No grouping or partial grouping results in group sizes of one with optimal machine utilizations that are balanced. Consequently, in both of these cases, $\rho_m = \rho, \ \forall m$.

In addition, for the partial grouping configuration, $L_{ij} = U_{ij} = 2$ in order to make it comparable to the total grouping case with equal group sizes. Preliminary experiments showed that for processing a particular operation, the policy of routing the part to the machine with the shorter queue performed consistently better than the policy in which the machine was selected randomly based upon probabilities that were specified *a priori.* Consequently, further experimentation considered partial grouping configurations only with state-dependent routing based on the shortest queue.

5.2.2 Loading Configurations

The alternative loading objectives discussed in §4 are used in conjunction with the grouping configurations mentioned in §5.2.1 to generate the system configurations listed in Table 3. C1 through C5 are constructed by solving **MLP** for the three cases of machine grouping. C1 is the base configuration, which is most similar to a conventional job shop. It is used primarily as a benchmark to evaluate the relative performance of the other configurations.

INSERT TABLE 3 HERE

Solving **MLPMT** results in configurations C6–C10, which parallel those obtained by solving Problem **MLP** under each machine grouping scenario. Configurations C11 and C12 are obtained by solving **MLPMF** for the (1, 2, 3) and (1, 1, 4) groupings. (Note that **MLPMF** applies only to these two grouping configurations because in all other cases, $\sum_m x_{ijm}$ is fixed: it equals 1 for no grouping, and equals 2 for partial grouping and for total grouping with (2, 2, 2) configurations.) Each of these two configurations is further decomposed into two subconfigurations corresponding to the way the weights are associated with the different operations. In C11A and C12A, all operations are given equal weights. This results in shorter operations being assigned to larger groups. Consequently, they will tend to be duplicated more often. In C11B and C12B, longer operations are assigned higher weights. This tends to generate configurations in which the longer operations are assigned to larger groups.

5.3 Experimental Results

The reported values of MFT are normalized with respect to the average part processing time $\overline{p}$. The first set of results corresponds to the impact of configurations C1—C5. These are discussed in §5.3.1. The configurations C6—C12 generated by considering the secondary loading objectives are dealt with in §5.3.2.

5.3.1 Impact of Grouping Configuration

Figure 5 shows the impact of CVOPT on MFT for these 5 configurations. Several results follow from these graphs. First, the performance of partial grouping and total grouping relative to no grouping improves with an increase in CVOPT. Second, partial grouping performs the best across all values of CVOPT and at all utilization levels. Once again, the relative superiority of using partial grouping increases with CVOPT; it also increases with an increase in the utilization level. Among the total grouping configurations, C3 is superior at low CVOPT; however, as CVOPT increases, the unbalanced configurations C4 and C5 perform better, especially at high utilizations. In particular, C5 is the best configuration at 90% utilization and at CVOPT=0.8.

INSERT FIGURE 5 HERE

These configurations exhibit varying levels of sensitivity to the scheduling rule used as shown in Table 4 for $\rho = 0.9$. (Results at other values of ρ are similar. Hence, they are not shown here.) C1 is most sensitive, and the impact of the scheduling rule increases with an increase in CVOPT. This is expected because as the difference among operation processing times increases, the impact of schedule quality increases. On the other hand C2 is, in general, least sensitive to the scheduling rule used. Its insensitivity does not depend upon CVOPT. C3–C5 show varying degrees of sensitivity, although, in all three cases, the impact of using a better scheduling rule increases with an increase in CVOPT. In particular, at CVOPT=0.8, the performance of C5 under SPT approaches that of C2.

INSERT TABLE 4 HERE

Table 5 depicts the impact of machine unreliability. First, note that in general, while increasing the level of unreliability increases MFT, the percentage increase comes down with an increase in CVOPT. This decrease is most prominent for the unbalanced configurations C4 and C5. Once again, C2 is the most robust configuration across all levels of unreliability, and its relative superiority improves with an increase in unreliability. Among the total grouping configurations, C3 is the most robust. As unreliability increases, it results in increasingly better values of MFT than both C4 and C5. Note that all groups have 2 machines in C3. Therefore, if one machine fails, an alternative machine is available to process parts. At the other extreme, C5 has two groups with one machine each. Hence, if any one of these machines fail, the operations of parts that are waiting at them are blocked.

INSERT TABLE 5 HERE

The impact of varying batch sizes is shown in Table 6. Once again we notice that as CVOPT increases, the adverse impact of larger batch sizes decreases. C2 remains the most effective configuration; however, its performance is closely matched by C3 at low CVOPT. As CVOPT increases, the relative performance of C3 deteriorates. Interestingly, the unbalanced configurations exhibit greater sensitivity to batch size, and they perform poorly as the batch sizes increase. For example, while C5 is superior to C3 at a batch size of 1 for CVOPT=0.8, the opposite is true when batch size increases to 10.

INSERT TABLE 6 HERE

5.3.2 Impact of Secondary Loading Objectives

Table 7 compares the performance of the configurations generated by solving **MLPMT** with those obtained from **MLP** at 0.9 system utilization level. The results show that, while operation aggregation has a mixed impact on MFT at low CVOPT, it leads to superior performance at high CVOPT values. This is partly explained by the fact that with an increase in CVOPT, the coefficient of variation of service times (C_s) at each machine increases; this, in turn, leads to a higher MFT. However, the aggregation of operations tends to reduce C_s, and therefore, MFT as well. Note, however, that in the case of partial grouping, operation aggregation is uniformly superior.

INSERT TABLE 7 HERE

Table 8 shows the impact of assigning operations based on the objective considered in **MLPMF**. Recall that C11A and C12A assign shorter operations to larger groups, and consequently provide greater flexibility to these operations. On the other hand, C11B and C12B provide greater flexibility to the longer operations. For comparison purposes, the MFT values obtained by solving Problem **MLP** are also included in Table 8. The results indicate that at low values of CVOPT, the C11B and C12B configurations are better, although they are comparable to C4 and C5, respectively. However, at higher CVOPT, C11A and C12A are significantly superior. This shows that at such high values of CVOPT, it is preferable to provide more alternative routes to as large a number of operations as possible. More importantly, this result shows that the weights that should be assigned to each operation to indicate its criticality are likely to depend upon CVOPT.

INSERT TABLE 8 HERE

6 Summary

This paper investigates the FMS planning problems of i) partitioning machines into groups, ii) determining the appropriate group utilization levels, and iii) assigning operations to these groups, for the objective of minimizing mean part flow time. Three grouping configurations — no grouping, partial grouping and total grouping, and three loading objectives are used for generating a variety of system configurations. An open queueing network representation of an FMS is used to show that, under total grouping, mean flow time is minimized when machine groups are maximally unbalanced, and the larger groups are utilized more heavily in terms of workload per machine. It is important to note that this result, as also the optimal group utilization levels derived in this study, hold when all machines are of the same type. If the FMS consists of several machine types, then in view of the findings of Dallery and Stecke (1990), it appears possible that the optimal grouping configuration for the entire system need not be the union of the optimal grouping configurations for each machine type considered individually. For the throughput function, Dallery and Stecke show that this condition is satisfied only when the optimal grouping configurations for each machine type are *N–dominant.* We are currently investigating equivalent conditions required for the mean flow time measure.

A simulation experiment is performed to study the impact of various system parameters on the performance of these configurations. Experimental results show that the importance of (partial or total) grouping increases with an increase in CVOPT. The relative merit of various configurations under total grouping depends upon, among other factors, the overall system utilization level and CVOPT. In particular, configurations with unequal group sizes are superior under high system utilizations and high processing time variations, and for more reliable systems with smaller fluctuations in production batch sizes.

Among the three grouping configurations considered, partial grouping with state-dependent routing is, in general, found to be superior across a range of different values that the various system parameters can take. Its performance is improved further if overall part movement is reduced by performing several operations of a part at the same machine. This improvement will increase when the actual travel times are accounted for. Aside from yielding superior flow time values, the partial grouping configuration is also robust in the face of machine failures and changes in production batch sizes, and is least sensitive to the quality of the scheduling rule employed.

Among the loading objectives, greater operation aggregation leads to superior performance at high CVOPT for all grouping configurations. For partial grouping and total grouping with unbalanced configurations, it does so at low CVOPT as well. Experimental results also indicate that when CVOPT is low, the assignment of longer operations should be duplicated more often. However, at high CVOPT, it is important to assign greater routing flexibility to a larger number of operations, by duplicating the shorter operations.

Acknowledgements

We thank the three anonymous reviewers and Avi Seidmann for their comments on an earlier version of this paper. Their remarks have helped to establish the scope of this work and improve the presentation of various results shown. Kathy Stecke would like to acknowledge a summer research grant from the Business School of The University of Michigan.

Appendix 1. Proof of Proposition 1

Proposition 1. *MFT does not decrease if any group g is decomposed into two subgroups $g1$ and $g2$, while the other groups remain unchanged.*

Proof. First, note that group g, and after decomposition subgroups $g1$ and $g2$, can be considered independent of the other groups. Let α_{g1} (α_{g2}), m_{g1} (m_{g2}), and λ_{g1} (λ_{g2}) denote, respectively, the visit ratio, number of machines and arrival rate for $g1$ ($g2$). Then, we have

$$m_g = m_{g1} + m_{g2}. \tag{31}$$

Also, from workload balance, we have

$$\frac{\lambda_g}{\mu_g} = \frac{\lambda_{g1}}{\mu_{g1}} + \frac{\lambda_{g2}}{\mu_{g2}}.$$

Hence,

$$m_g\rho_g = m_{g1}\rho_{g1} + m_{g2}\rho_{g2}. \tag{32}$$

The increase in MFT because of decomposing g is

$$\begin{aligned}\Delta MFT &= [(\alpha_{g1}W_{g1} + \alpha_{g2}W_{g2}) - \alpha_g W_g] \\ &= \frac{1}{\lambda}[(L_{g1} + L_{g2}) - L_g].\end{aligned}$$

For the trivial case in which m_{g1} or m_{g2} equals zero, clearly $\Delta MFT = 0$. Otherwise note that

$$\begin{aligned}L_g &= \frac{\lambda_g}{\mu_g} + L_g^q \\ &= m_g\rho_g + L_g^q,\end{aligned}$$

where L_g^q denotes the mean queue length (parts waiting for service) at group g. Hence,

$$\begin{aligned}\Delta MFT &= \frac{1}{\lambda}[\{(m_{g1}\rho_{g1} + m_{g2}\rho_{g2}) - m_g\rho_g\} + \{(L_{g1}^q + L_{g2}^q) - L_g^q\}] \\ &= \frac{1}{\lambda}[(L_{g1}^q + L_{g2}^q) - L_g^q].\end{aligned}$$

From queueing theory, we know that the mean queue length L_c^q in a single-channel system with c parallel servers is given by

$$L_c^q = p_0\frac{(c\rho)^c\rho}{c!(1-\rho)^2}$$

where ρ is the server utilization, and p_0 is the probability that an arriving part finds the system empty. Note that L_c^q is convex in ρ for a given c, and convex in c for a given ρ.

From Equations (31) and (32), it follows that ρ_g is a convex combination of ρ_{g1} and ρ_{g2}. Therefore, $L_g^q < L_{g1}^q + L_{g2}^q$, and consequently $\Delta MFT > 0$.

This proves the proposition. □

Appendix 2. MFT under the (1, 2) configuration

The MGP for the (1, 2) configuration can be written as

$$\min \ MFT = \frac{1}{\lambda}\left[\frac{\rho_1}{1-\rho_1} + \frac{2\rho_2}{1-{\rho_2}^2}\right]$$

subject to

$$\rho_1 + 2\rho_2 = 3\rho \tag{33}$$

$$0 \le \rho_g < 1, \ \ g = 1 \ \ and \ \ 2.$$

Associating the multiplier u with constraint (33) and using the Kuhn-Tucker conditions yields the following relationships at the optimal solution to minimize MFT:

$$u = \frac{1}{\lambda(1-\rho_1)^2} = \frac{1+\rho_2^2}{\lambda(1-\rho_2^2)^2}$$

or

$$\rho_1 = 1 - \frac{1-\rho_2^2}{\sqrt{1+\rho_2^2}} \tag{34}$$

From Equations (33) and (34), we have

$$(4-12\rho)\rho_2 + (7-6\rho+9\rho^2)(\rho_2)^2 + (4-12\rho)(\rho_2)^3 + 3(\rho_2)^4 - 6\rho + 9\rho^2 = 0,$$

and

$$\rho_1 = 3\rho - 2\rho_2.$$

References

1. Ammons, J. C., C. B. Lofgren and L. F. McGinnis (1984), "A Large Scale Workstation Loading Problem," in *Proceedings of the First ORSA/TIMS Conference on Flexible Manufacturing Systems*, Ann Arbor, MI, 249–255.

2. Arbib, C., M. Lucertini and F. Nicolò (1991), "Workload Balance and Part-Transfer Minimization in Flexible Manufacturing Systems," *International Journal of Flexible Manufacturing Systems*, Vol. 3, 5–25.

3. Berrada, M. and K. E. Stecke (1986), "A Branch and Bound Approach for Machine Load Balancing in Flexible Manufacturing Systems," *Management Science*, Vol. 32, 1316–1335.

4. Buzacott, J. A. and D. D. Yao (1986), "Flexible Manufacturing Systems: A Review of Analytical Models," *Management Science*, Vol. 32, 890–905.

5. Chow, Y.-C. and W. H. Kohler (1979), " Models for Dynamic Load Balancing in a Heterogeneous Multiple Processor System," *IEEE Transactions on Computers*, Vol. C-28, 354–361.

6. Conway, R. W., W. L. Maxwell and L. W. Miller (1967), *Theory of Scheduling*, Addison-Wesley, Reading, MA.

7. Dallery, Y. and K. E. Stecke (1990), "On the Optimal Allocation of Servers and Workloads in Closed Queueing Networks," *Operations Research*, Vol. 38, 694–703.

8. Hwang, S. (1986), "A Constraint-Directed Method to Solve the Part Selection Problem in Flexible Manufacturing Systems Planning Stage", in *Proceedings of the Second ORSA/TIMS Conference on Flexible Manufacturing Systems*, Ann Arbor, MI, Elsevier Science Publishers B.V., Amsterdam, 297–309.

9. Kleinrock, L. (1975), *Queueing Systems: Volume 1: Theory*, John Wiley and Sons, New York, NY.

10. Kochman, G. A. (1989), "The Performance of Manufacturing Systems in a Disruptive Environment," presented at the TIMS Meeting, Osaka, Japan.

11. Kusiak, A. (1984), "The Parts Families Problem in Flexible Manufacturing Systems," in *Proceedings of the First ORSA/TIMS Conference on Flexible Manufacturing Systems*, Ann Arbor, MI, 237–242.

12. Lee, H. L. and M. A. Cohen (1983), "A Note on the Convexity of Performance Measures of $M/M/c$ Queueing Systems," *Journal of Applied Probability*, Vol. 20, 920–923.

13. Monahan, G. E. and T. L. Smunt (1990), "Product-Process Relations in Batch Manufacturing," Working Paper # 90–1702, College of Commerce and Business Administration, University of Illinois at Urbana-Champaign.

14. Rajagopalan, S. (1986), "Formulation and Heuristic Solutions for Parts Grouping and Tool Loading in Flexible Manufacturing Systems", in *Proceedings of the Second ORSA/TIMS Conference on Flexible Manufacturing Systems*, Ann Arbor, MI, Elsevier Science Publishers B.V., Amsterdam, 311-320.

15. Sauer, C. H. (1983), "Computational Algorithms for State-Dependent Queueing Networks," *ACM Transactions on Computer Systems*, Vol. 1, 67–92.

16. Shanthikumar, J. G. and D. D. Yao (1987), "Optimal Server Allocation in a System of Multi-Server Stations," *Management Science*, Vol. 33, 1173–1180.

17. Shanthikumar, J. G. and D. D. Yao (1988), "On Server Allocation in Multiple Center Manufacturing Systems," *Operations Research*, Vol. 36, 333–342.

18. Stecke, K. E. (1983), "Formulation and Solution of Nonlinear Integer Production Planning Problems for Flexible Manufacturing Systems," *Management Science*, Vol. 29, 273–288.

19. Stecke, K. E. (1986a), "On the Nonconcavity of Throughput in Certain Closed Queueing Networks," *Performance Evaluation*, Vol. 6, 293–305.

20. Stecke, K. E. (1986b), "A Hierarchical Approach to Solving Machine Grouping and Loading Problems of Flexible Manufacturing Systems," *European Journal of Operational Research*, Vol. 24, 369–378.

21. Stecke, K. E. and I. Kim (1989), "Performance Evaluation for Systems of Pooled Machines of Unequal Sizes: Unbalancing Versus Balancing," *European Journal of Operational Research*, Vol. 42, 22–38.

22. Stecke, K. E. and I. Kim (1991), "A Flexible Approach to Part Type Selection in Flexible Flow Systems Using Part Mix Ratios," *International Journal of Production Research*, Vol. 29, 53–75.

23. Stecke, K. E. and J. J. Solberg (1985), "The Optimality of Unbalancing Both Workloads and Machine Group Sizes in Closed Queueing Networks of Multiserver Queues," *Operations Research*, Vol. 33, 882–910.

24. Towsley, D. F. (1980), "Queueing Network Models with State-Dependent Routing," *Journal of the Association for Computing Machinery*, Vol. 27, 323–337.

25. Winston, W. (1977), "Optimality of the Shortest Line Discipline," *Journal of Applied Probability*, Vol. 14, 181–189.

26. Yao, D. D. and J. A. Buzacott (1985), "Modeling a Class of State-Dependent Routing in Flexible Manufacturing Systems," *Annals of Operations Research*, Vol. 3, 153–167.

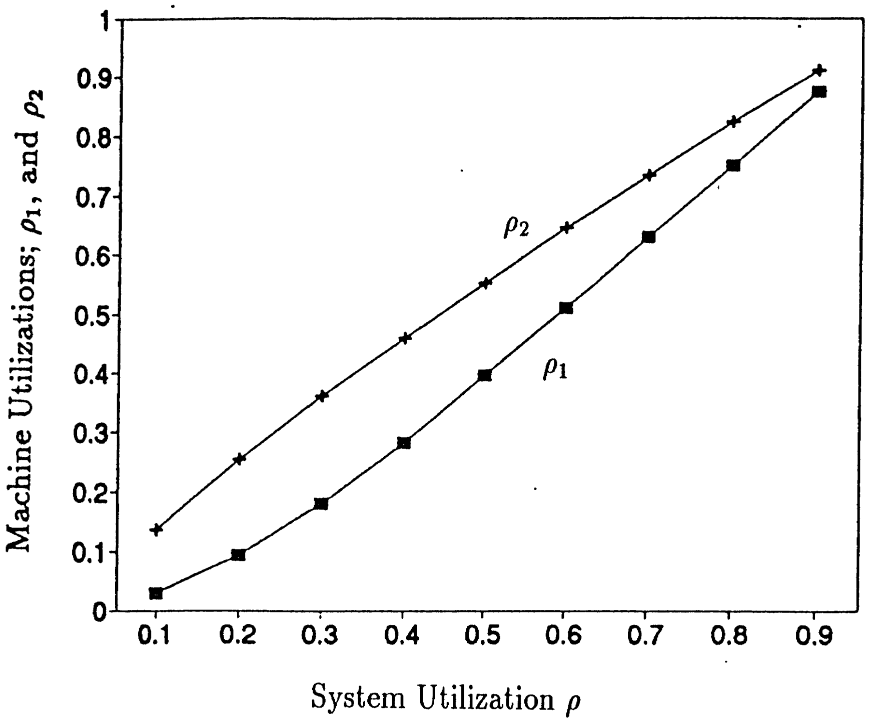


Figure 1: Optimal Utilization per Machine within Each Group: 3 machines, 2 groups

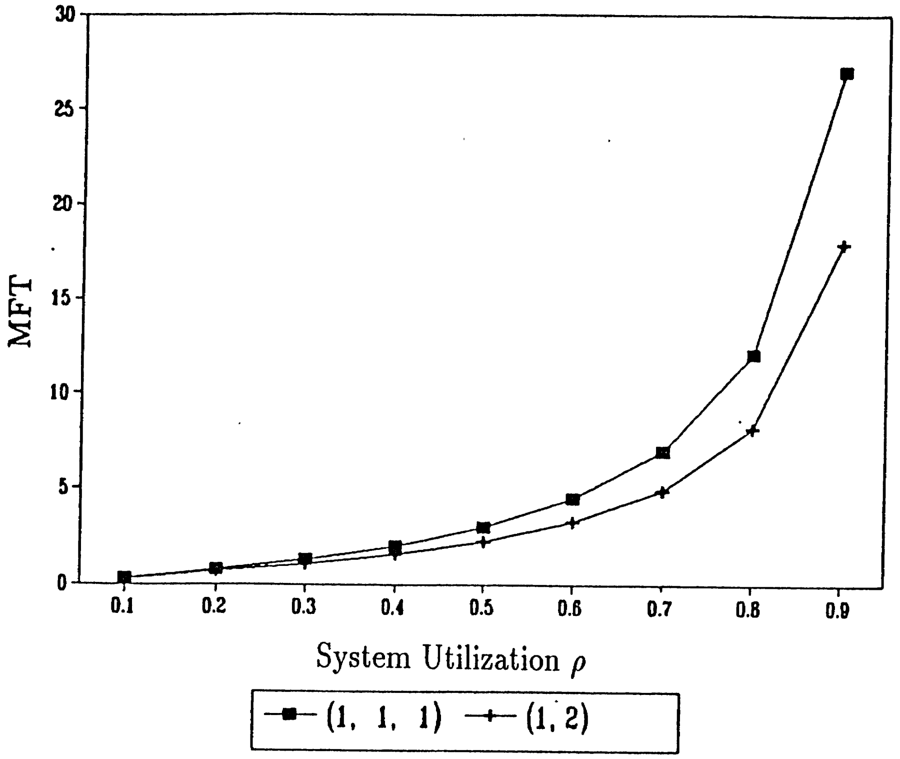


Figure 2: MFT under Alternative Grouping Configurations: 3 machines, $\lambda = 1$

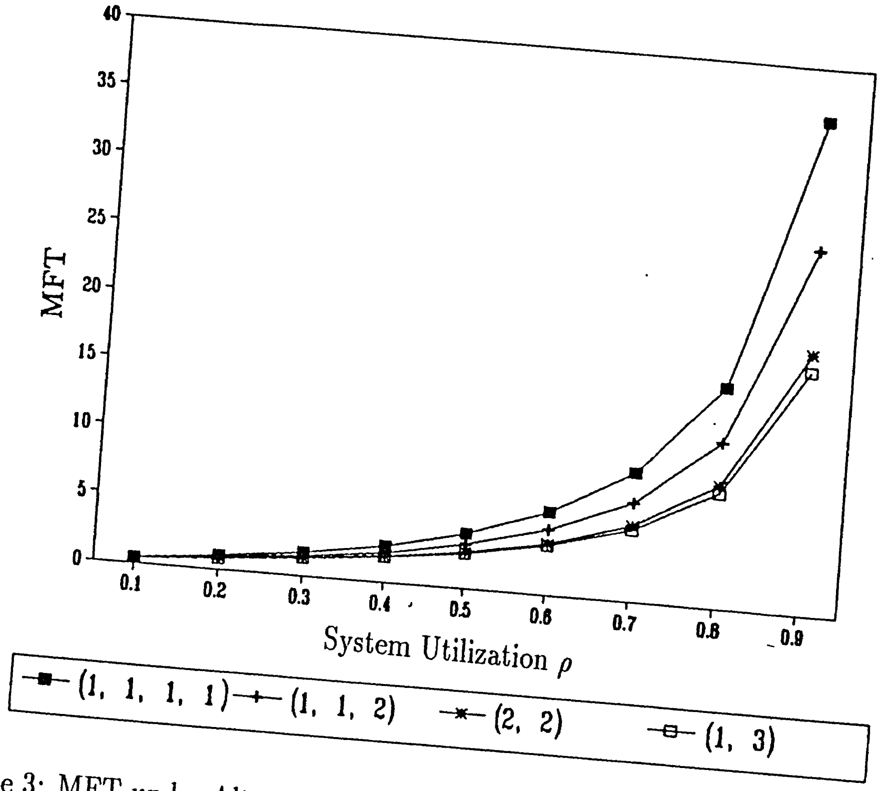


Figure 3: MFT under Alternative Grouping Configurations: 4 machines, $\lambda = 1$

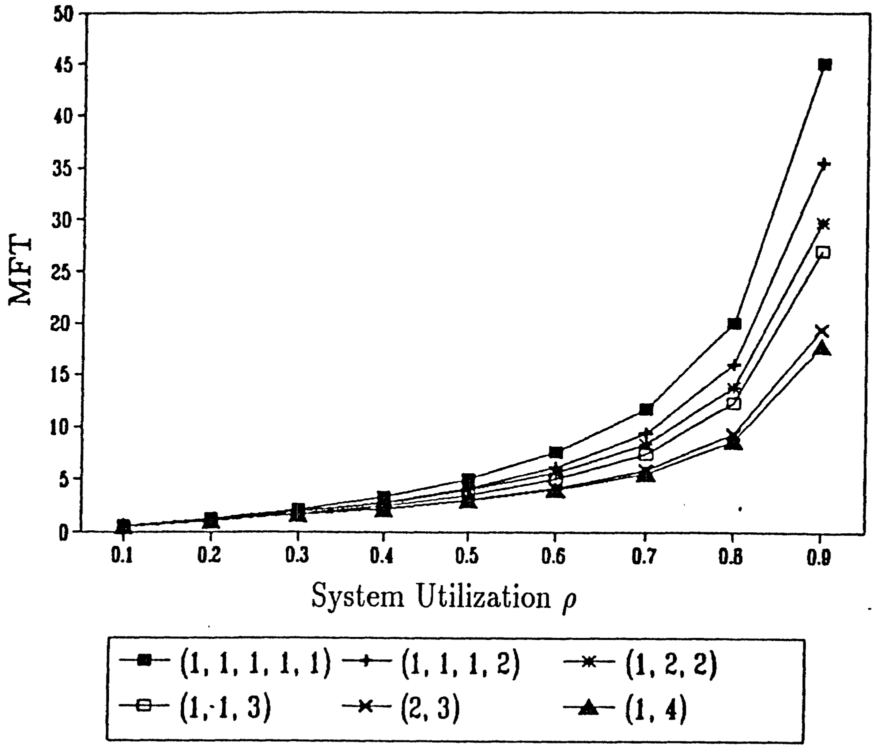
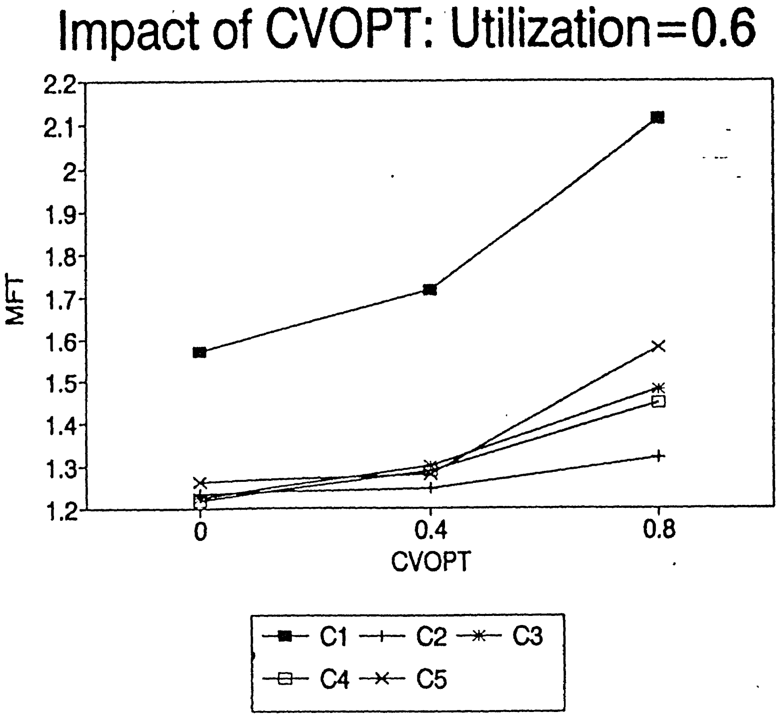


Figure 4: MFT under Alternative Grouping Configurations: 5 machines, $\lambda = 1$



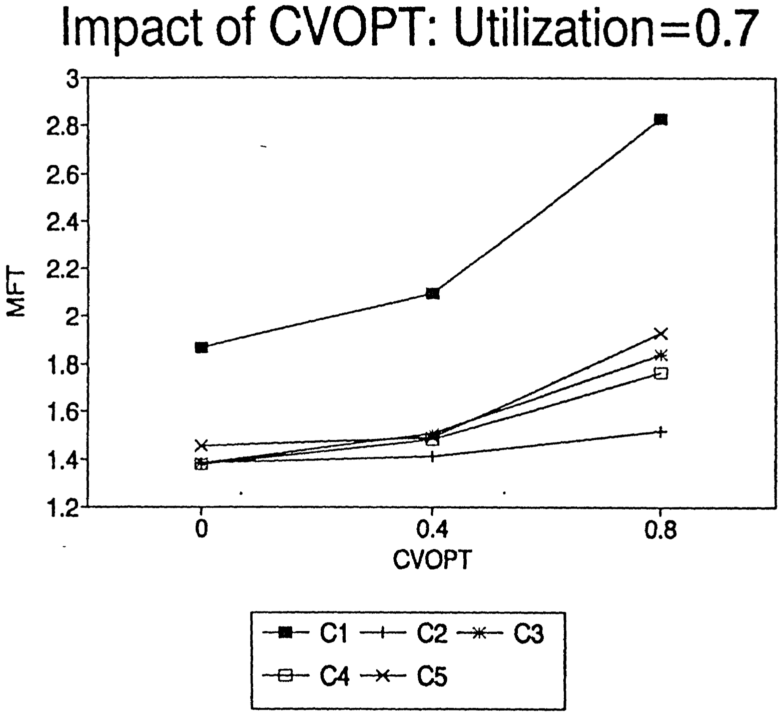
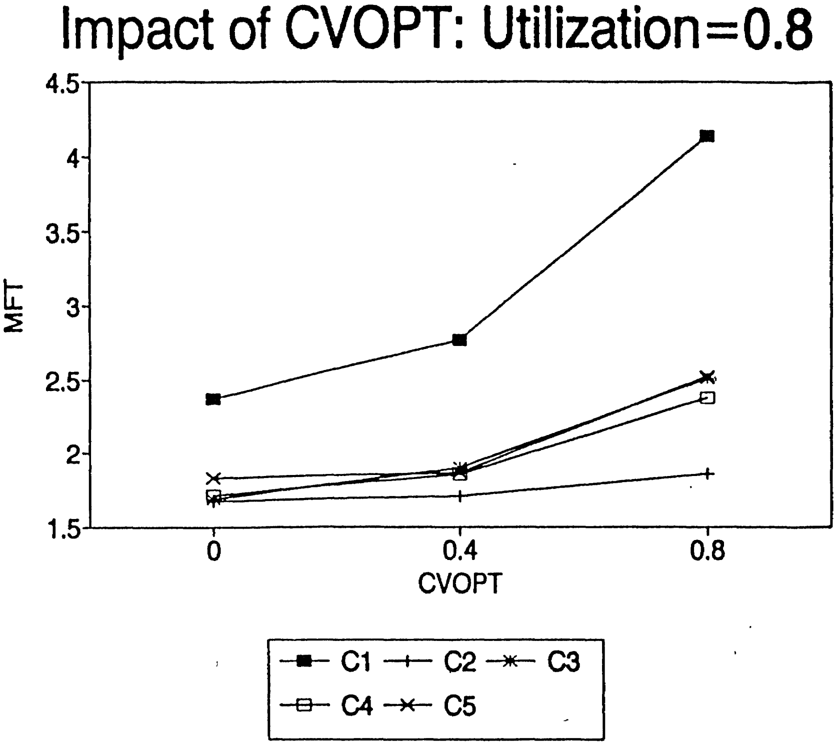


Figure 5: Impact of CVOPT on MFT



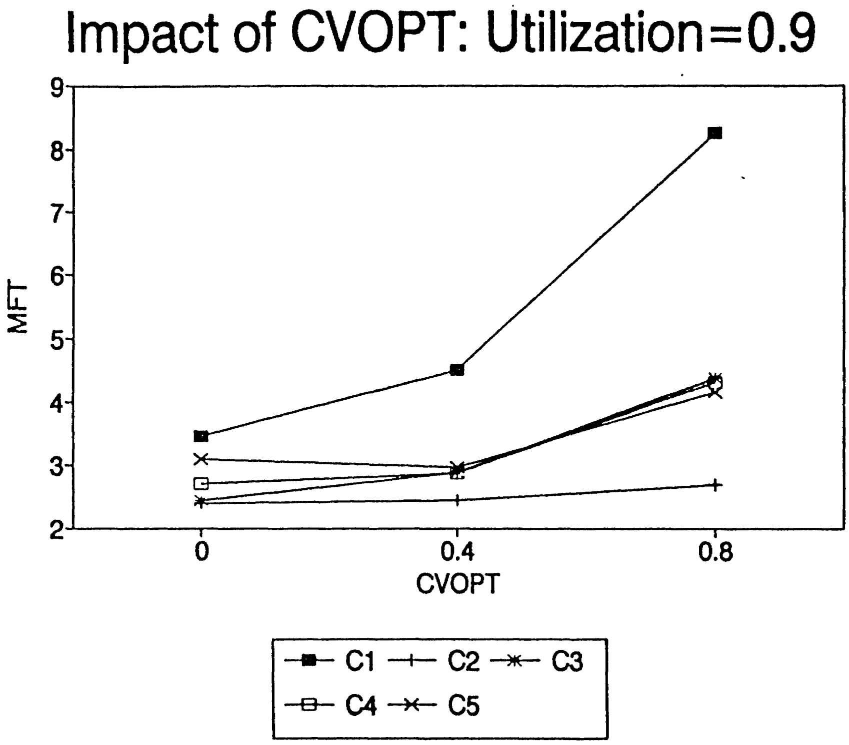


Figure 5 (continued): Impact of CVOPT on MFT

TABLE 1
Notation

Parameters

n_j Number of operations for a part of type j, $j = 1, 2, \ldots, N$.

p_{ij} Processing time of operation i of a part of type j,
$j = 1, 2, \ldots, N; i = 1, 2, \ldots, n_j$.

t_{ij} Number of tool slots required for operation i of part type j,
$j = 1, 2, \ldots, N; i = 1, 2, \ldots, n_j$.

T^m Tool magazine capacity of machine m, $m = 1, 2, \ldots, M$.

λ^j Arrival rate of parts of type j, $j = 1, 2, \ldots, N$.

λ Cumulative arrival rate of all parts at the system $= \sum_{j=1}^{N} \lambda^j$.

λ_{ijm} Arrival rate of operation i of part type j at machine m.

B A large number.

Variables

ρ_m Utilization (workload) of machine $m, m = 1, 2, \ldots, M$.

W_{ijm} (Steady-state) average time spent by a part of type j at machine m
for operation i, $j = 1, 2, \ldots, N; i = 1, 2, \ldots, n_j; m = 1, 2, \ldots, M$.

x_{ijm} Operation to machine assignment variable,
$x_{ijm} = 1$, if operation i of part type j is assigned to machine m; 0, otherwise.

TABLE 2

Optimal Utilization per Machine within Each Group – Total Grouping

	(1, 2, 3)			*(1, 1, 4)*			*(2, 2, 2)*
ρ^*	ρ_1^*	ρ_2^*	ρ_3^*	ρ_1^*	ρ_2^*	ρ_3^*	$\rho_1^* = \rho_2^* = \rho_3^*$
0.1	0.008	0.074	0.148	0.002	0.002	0.149	0.1
0.2	0.044	0.173	0.270	0.026	0.026	0.287	0.2
0.3	0.111	0.276	0.378	0.084	0.084	0.408	0.3
0.4	0.203	0.382	0.478	0.174	0.174	0.513	0.4
0.5	0.313	0.487	0.571	0.288	0.288	0.606	0.5
0.6	0.438	0.589	0.661	0.418	0.418	0.691	0.6
0.7	0.572	0.694	0.747	0.558	0.558	0.771	0.7
0.8	0.712	0.796	0.832	0.702	0.702	0.849	0.8
0.9	0.855	0.890	0.917	0.850	0.850	0.925	0.9

TABLE 3

System Configurations Examined with Various Loading Objectives

System Configuration	*Loading Objective*	*Level of Grouping*	*Grouping Configuration*
C1	MLP	No Grouping	(1, 1, 1, 1, 1, 1)
C2	MLP	Partial Grouping	(1, 1, 1, 1, 1, 1)
C3	MLP	Total Grouping	(2, 2, 2)
C4	MLP	Total Grouping	(1, 2, 3)
C5	MLP	Total Grouping	(1, 1, 4)
C6	MLPMT	No Grouping	(1, 1, 1, 1, 1, 1)
C7	MLPMT	Partial Grouping	(1, 1, 1, 1, 1, 1)
C8	MLPMT	Total Grouping	(2, 2, 2)
C9	MLPMT	Total Grouping	(1, 2, 3)
C10	MLPMT	Total Grouping	(1, 1, 4)
C11 (A, B)	MLPMF	Total Grouping	(1, 2, 3)
C12 (A, B)	MLPMF	Total Grouping	(1, 1, 4)

TABLE 4

Impact of Scheduling Rules on MFT at $\rho = 0.9$

CVOPT	*Configuration*	*MFT under*		*% Decrease under SPT*
		FCFS	SPT	
0.0	C1	3.46	3.46	0.0
	C2	2.40	2.40	0.0
	C3	2.45	2.45	0.0
	C4	2.72	2.72	0.0
	C5	3.11	3.11	0.0
0.4	C1	4.50	3.49	22.6
	C2	2.44	2.29	6.3
	C3	2.90	2.57	11.4
	C4	2.88	2.68	7.1
	C5	2.97	2.89	2.7
0.8	C1	8.25	4.75	42.5
	C2	2.69	2.52	6.0
	C3	4.38	3.33	23.9
	C4	4.31	3.04	28.8
	C5	4.14	2.77	33.2

TABLE 5

Impact of Machine Breakdowns on MFT at $\rho = 0.9$

CVOPT	*Configuration*	*MFT under Level*			*% Increase over Level 0*	
		0	1	2	Level 1	Level 2
0.0	C1	3.46	4.61	6.41	33.2	85.4
	C2	2.40	2.80	3.37	16.5	40.3
	C3	2.45	3.41	4.94	39.3	101.7
	C4	2.72	4.94	6.86	81.9	152.4
	C5	3.11	5.04	9.38	62.2	201.6
0.4	C1	4.50	5.91	8.40	31.2	86.5
	C2	2.44	2.86	3.40	16.9	38.8
	C3	2.90	3.95	5.56	36.5	92.1
	C4	2.88	4.10	6.52	42.2	126.6
	C5	2.97	4.38	6.94	47.7	133.5
0.8	C1	8.25	10.77	13.92	30.5	68.6
	C2	2.69	3.08	3.58	14.9	33.2
	C3	4.38	5.64	7.70	28.7	75.9
	C4	4.31	5.92	8.31	37.5	92.8
	C5	4.14	5.78	8.89	39.5	146.9

TABLE 6

Impact of Batch Size on MFT at $\rho = 0.9$

CVOPT	*Configuration*	*MFT under Batch Size*			*% Increase over BS = 1*	
		1	5	10	BS = 5	BS = 10
0.0	C1	3.46	8.29	13.92	139.6	302.4
	C2	2.40	7.54	13.50	214.1	462.5
	C3	2.45	7.57	13.51	208.8	451.5
	C4	2.72	9.20	15.86	238.6	483.9
	C5	3.11	11.18	18.75	259.8	503.1
0.4	C1	4.50	9.77	15.31	116.8	239.9
	C2	2.44	7.55	13.46	208.8	450.4
	C3	2.90	8.02	13.70	176.8	372.9
	C4	2.88	8.88	15.19	208.5	427.6
	C5	2.97	8.90	15.31	199.7	415.3
0.8	C1	8.25	14.27	19.83	72.9	140.2
	C2	2.69	7.79	13.80	189.9	413.3
	C3	4.38	10.00	16.10	128.4	267.6
	C4	4.31	10.87	17.72	152.3	311.4
	C5	4.14	10.24	16.58	147.4	300.3

TABLE 7

Impact of Operation Aggregation on MFT at $\rho = 0.9$

Configuration	*CVOPT*		
	0.0	0.4	0.8
No Grouping			
C1	3.46	4.50	8.25
C6	4.75	5.60	5.41
Partial Grouping			
C2	2.40	2.44	2.69
C7	2.25	2.31	2.33
Total Grouping			
C3	2.45	2.90	4.38
C8	2.85	3.10	3.79
C4	2.72	2.88	4.38
C9	2.70	2.75	2.77
C5	3.11	2.97	4.14
C10	2.97	3.02	3.11

TABLE 8

Impact of Operation Duplication on MFT at $\rho = 0.9$

Configuration	*CVOPT*	
	0.4	0.8
C4	2.88	4.31
C11A	3.39	3.52
C11B	3.20	3.74
C5	2.97	4.14
C12A	3.28	3.46
C12B	2.91	4.62