division of research
GRADUATE SCHOOL OF BUSINESS ADMINISTRATION
UNIVERSITY OF MICHIGAN ANN ARBOR, MICHIGAN 48109

REDUCING WORK-IN-PROCESS INVENTORY IN CERTAIN CLASSES OF FLEXIBLE MANUFACTURING SYSTEMS

Working Paper No. 407

J. George Shanthikumar
School of Business Administration
University of California, Berkeley
and
Kathryn E. Stecke
Graduate School of Business Administration
The University of Michigan

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

December 1984

REDUCING WORK-IN-PROCESS INVENTORY IN CERTAIN CLASSES OF FLEXIBLE MANUFACTURING SYSTEMS

Working Paper No. 407

J. George Shanthikumar
School of Business Administration
University of California, Berkeley
and
Kathryn E. Stecke
Graduate School of Business Administration
The University of Michigan

Abstract

In this paper, we establish that maintaining a balanced workload on each machine over time stochastically minimizes the work-in-process inventory in certain types of flexible manufacturing systems (FMSs) with finite or infinite common input buffer storage and an ample buffer at each machine. The results obtained here complement those obtained by Stecke and Morin (1984), in which it is established that balancing workloads maximizes expected production, again for the same, particular types of FMSs. Stecke and Morin (1984) treats a static FMS loading problem, while this paper addresses a dynamic problem which considers three strategies to control the release of parts into the system.

Key words: Flexible manufacturing systems, work-in-process inventory, workload balancing, stochastic minimization.

1. INTRODUCTION

A flexible manufacturing system (FMS) is an automated alternative to the conventional means of batch manufacturing, to date applied mainly in the metal-cutting industry. An FMS consists of a number of computer numerically controlled machine tools, which are linked together by an automated material handling system. Individual parts of different types can be machined simultaneously in unit batch sizes.

Efficient use of such FMSs requires careful pre-production planning (see, e.g., Buzacott and Shanthikumar (1980) and Stecke (1981)). One of such planning problems concerns the appropriate loading of the machines in an FMS. By FMS machine loading, we mean the allocation of all operations and their associated cutting tools among the FMS machine tools, in order to define which machines will be able to perform each operation of the part types that have been chosen to be machined next.

Earlier analyses of this and similar planning issues in the context of job shops and flow lines have indicated the superiority of either unbalancing or balancing each machine's workload (depending on some particular aspects of the production system, such as buffer size, processing time distribution,...) with respect to the maximization of production rate or with respect to the minimization of the work-in-process inventory. (See Stecke and Morin (1982) for references to such works.)

The question of balancing workloads in the context of FMS was first raised by Buzacott and Shanthikumar (1980). Based on an asymptotic analysis, they established that balancing the workload on all machines maximizes the expected production rates whenever the number of parts in the system is very large, i.e., approaching infinity. Subsequently, Stecke (1981) and Stecke and Morin (1982), based on empirical results, supported the optimality of balanced

workloads with respect to the expected production of FMSs with single machine workcenters (i.e., with no pooling of similar machines) and with a finite number of parts.

However, it was also observed in Stecke (1981) that balancing the workload per machine need not maximize the expected production of FMSs having a different number of machines in different machine centers (i.e., with pooling of machines). Shanthikumar (1982) for the first time provided an analytical proof for the optimality of the balanced load with respect to the production rate of an FMS with no pooling. An empirical proof can be found in Stecke (1981). An alternative analytical proof can be found in Stecke and Morin (1984). Several extensions and simpler proofs of similar results can be found in Yao (1984a, b) and Yao and Kim (1984). All of these FMS loading problems involve static allocations of workload among machines (and eventually, of operations and cutting tools). These problems are solved and the solutions implemented before the FMS begins to operate. The detailed, actual FMS loading problem of allo-allocating operations and tools to machines to maximize production has been addressed in Stecke (1983, 1985). Again, the problem is static.

The scenario in this paper is different, but complementary to the previous studies, Stecke and Morin (1984), in particular. An open queueing network is used to model the random arrival of individual parts to the FMS. Given that there are N parts in the system, a closed queueing network then provides the expected production, if the workload on each machine is also provided. Three strategies are considered to release the parts, which have arrived into a production control area, into the FMS. Under these three release policies, if the workload per machine remains balanced over time, not only is the expected production maximized, but we show that the in-process inventory is stochastically minimized.

The random arrivals are of individual parts, and not batches, of a limited number of similar part types. Each operation can be performed on only one machine. The cutting tools for each operation have already been loaded into the appropriate tool magazine. There is no pooling of machines and there is a finite or infinite common buffer area (called a production control area) for incoming parts, which are then released into the FMS. This paper is treating this dynamic problem. The solution is that it is optimal under several objectives to maintain a balanced load on each machine over time as the parts are input into the FMS.

The open and closed queueing network models are described in §2, as well as the three policies that we consider to release parts into the FMS. The optimality of balancing the workload on each machine to maximize expected production using the closed queueing network is presented in §3. §4 contains the stochastic minimization of the number of parts in the FMS. A summary and future research is provided in §5.

2. THE QUEUEING NETWORK MODELS

A Jackson type queueing network is used to model a flexible manufacturing system. The external arrival of parts forms a Poisson process with rate λ. An external arrival is received into a production control area from which the parts are dispatched into the flexible manufacturing system. (See Figure 1.) Parts from the production control area are released to the flexible manufacturing system according to some dispatch policy. In this paper, we consider the following three policies that describe different means to control the input of parts to the FMS:

Policy I: Dispatch parts to the FMS as soon as they are received into the production control area.

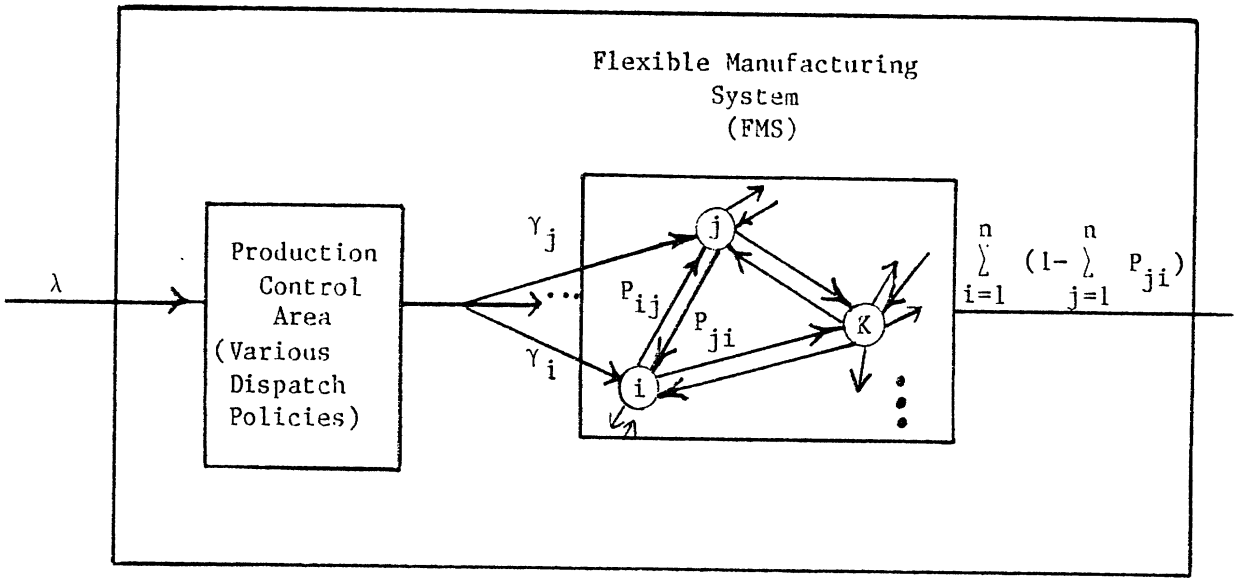

Figure 1: A Schematic Diagram of a Flexible Manufacturing System.

Policy II: Any parts that are received into the production control area when there are already Z parts in the FMS are rejected (i.e., lost). Otherwise, parts are released to the FMS as soon as they are received by the production control area.

Policy III: Dispatch a part to the FMS if and only if the total number of parts in the FMS is less than (some value) Z. This may represent a predetermined input control policy or reflect the fact that the number of pallets available in the shop is limited to Z (e.g., see Buzacott (1982) and Buzacott and Shanthikumar (1980)).

The flexible manufacturing system contains M machines (each is single, there is no pooling). The first operation of a part that is released to the FMS is performed by machine i with probability γ_i. A part that has completed its processing at machine i will proceed next to machine j with probability

p_{ij}. This part is an internal arrival to machine j. On the other hand, upon leaving machine i, a part may depart the network of queues (i.e., the FMS) with probability $1 - \sum_{j=1}^{M} p_{ij}$. The service, or processing, time at machine i is exponentially distributed with mean t_i, i=1,2,...,M. Let q_i be the expected number of visits made to machine i by an arbitrary part. Then, using the Markovian property of the part transitions from machine to machine, it can be shown that:

$$q_i = \gamma_i + \sum_{j=1}^{M} q_j p_{ji}, \quad i=1,2,\ldots,M. \tag{1}$$

The mean load imposed on machine i by a single part is then equal to $q_i t_i$ and we denote this by x_i. If λ_e is the effective part arrival rate to the FMS (different for the different part input policies), then the workload rate utilization of machine i is $\lambda_e x_i$. The total mean workload L imposed on the system by a single part is equal to $\sum_{i=1}^{M} x_i$. L is the average time required to process a part through the system. We will assume that, independent of how the machines are loaded, the mean total load will remain the constant L.

Then, under a balanced workload, the mean load imposed by a single part on a machine tool is L/M ($\equiv x^*$). The balanced workload distribution is then:

$$\underline{x}^* = (x^*, x^*, \ldots, x^*).$$

3. PRODUCTION RATE

In this section, the number of parts in the system is kept constant at level N, i.e., the queueing network is closed. Then the expected production rate, $PR_N(\underline{x})$, of the FMS with the mean load distribution $\underline{x} = (x_1, x_2, .., x_M)$ is (see, e.g., Solberg (1977) and Buzacott and Shanthikumar (1980)),

$$PR_N(\underline{x}) = g(\underline{x}, N-1)/g(\underline{x}, N), \quad N \geq 1, \tag{2}$$

where

$$g(\underline{x},N) = \sum_{\underline{n}\in S_N} \left(\prod_{i=1}^{M} x_i^{n_i} \right), \quad N \geq 1 \tag{3}$$

$$g(\underline{x},0) = 1,$$

$$\underline{n} = (n_1,n_2,..,n_M), \quad n_i \geq 0, \quad i=1,2,..,M,$$

and

$$S_N = \{\underline{n}: \sum_{i=1}^{M} n_i = N\}, \quad N \geq 1.$$

Here, n_i is the number of parts at machine i, including the part in process if $n_i \geq 1$.

The following Theorem concerning maximum expected production, using different proofs, has been established in Stecke (1981), Shanthikumar (1982), Stecke and Morin (1982), and Yao (1984b).

Theorem 1: The expected production rate, $PR_N(\underline{x})$, is maximized by balancing the workload on all machines. That is,

$$PR_N(\underline{x}^*) = \max_{\underline{x}: \sum_{i=1}^{M} x_i = L} \{PR_N(\underline{x})\}$$

for $\underline{x}^* = (L/M, L/M, \ldots, L/M)$.

Corollary 2: The expected production rate of a single-server closed queueing network, $PR_N(\underline{x})$, is bounded by:

$$PR_N(\underline{x}) \leq \frac{N}{M+N-1} \frac{M}{L}. \tag{4}$$

We note here that for information concerning the robustness of queueing networks to aggregately model the steady state performance of FMSs, see Suri (1983), for example.

4. NUMBER OF JOBS IN THE SYSTEM

Let $P_n(\underline{x})$ be the steady state probability that there are n parts (n=0,1,...) in the system at an arbitrary time epoch under a mean work load distribution $\underline{x} = (x_1, x_2, \ldots, x_M)$. It can be shown (see, e.g., Shanthikumar and Sargent (1981)) that under Policy I:

$$\left.\begin{aligned} P_n(\underline{x}) &= \frac{\lambda}{PR_n(\underline{x})} P_{n-1}(\underline{x}), \quad n=1,2,\ldots; \\ P_0(\underline{x}) &= \left[1 + \sum_{n=1}^{\infty} \left\{ \prod_{i=1}^{n} \left(\frac{\lambda}{PR_i(\underline{x})}\right)\right\}\right]^{-1}. \end{aligned}\right\} \tag{5}$$

and under Policy II:

$$\left.\begin{aligned} P_n(\underline{x}) &= \frac{\lambda}{PR_n(\underline{x})} P_{n-1}(\underline{x}), \quad n=1,2,\ldots,Z; \\ P_n(\underline{x}) &= 0, \quad n=Z+1,Z+2\ldots; \\ P_0(\underline{x}) &= \left[1 + \sum_{n=1}^{Z} \left\{ \prod_{i=1}^{n} \left(\frac{\lambda}{PR_i(\underline{x})}\right)\right\}\right]^{-1}. \end{aligned}\right\} \tag{6}$$

For Policy III, Buzacott and Shanthikumar (1980) has established that the following is a good approximation.

$$\left.\begin{aligned} P_n(\underline{x}) &= \frac{\lambda}{PR_n(\underline{x})} P_{n-1}(\underline{x}), \quad n=1,2,\ldots,Z; \\ P_n(\underline{x}) &= \frac{\lambda}{PR_Z(\underline{x})} P_{n-1}(\underline{x}), \quad n=Z+1,Z+2,\ldots; \\ P_0(\underline{x}) &= \left[1 + \sum_{n=1}^{Z-1} \left\{ \sum_{i=1}^{n} \left(\frac{\lambda}{PR_i(\underline{x})}\right)\right\} + \prod_{i=1}^{Z} \frac{\lambda}{PR_i(\underline{x})} \cdot \left(\frac{PR_Z(\underline{x})}{PR_Z(\underline{x})-\lambda}\right)\right]^{-1} \end{aligned}\right\} \tag{7}$$

Let M(n)/M(n)/1 be a birth-death queueing system with state-dependent arrival rates $\underline{\lambda} = (\lambda_n)_0^\infty$ and state-dependent service rates $\underline{\mu} = (\mu_n)_1^\infty$. Also, let $F_n(\underline{\lambda},\underline{\mu})$ be the steady state probability that the number of parts in the system is less than or equal to n. Then:

$$F_n(\underline{\lambda},\underline{\mu}) = \sum_{r=0}^{n} \rho_r(\underline{\lambda},\underline{\mu}) \; / \; \sum_{r=0}^{\infty} \rho_r(\underline{\lambda},\underline{\mu}), \tag{8}$$

where

$$\left.\begin{aligned} \rho_0(\underline{\lambda},\underline{\mu}) &= 1 \\ \rho_r(\underline{\lambda},\underline{\mu}) &= \prod_{i=1}^{r} \frac{\lambda_{i-1}}{\mu_i}, \quad r=1,2,\ldots \end{aligned}\right\} \tag{9}$$

The results (5), (6), and (7) are special cases of (8) with:

- $\lambda_n = \lambda$ (n=0,1,...) and $\mu_n = PR_n(\underline{x})$ (n=1,2,...) for Policy I;
- $\lambda_n = \lambda$ (n=0,1,...,Z-1), $\lambda_n = 0$ (n=Z,Z+1,...), and $\mu_n = PR_n(x)$ (n=1,2,...) for Policy II;
- and $\lambda_n = \lambda$ (n=0,1,...), $\mu_n = PR_n(\underline{x})$ (n=1,2,...,Z), and $\mu_n = PR_Z(\underline{x})$ (n=Z+1,Z+2,...) for Policy III.

Now, taking the partial derivatives of (8) with respect to λ_r and μ_r, one can see that:

$$\left.\begin{aligned} &\frac{\partial}{\partial\lambda_r} F_n(\underline{\lambda},\underline{\mu}) \leq 0, \quad r = 0,1,\ldots \\ \text{and} \quad &\frac{\partial}{\partial\mu_r} F_n(\underline{\lambda},\underline{\mu}) \geq 0, \quad r = 1,2,\ldots \end{aligned}\right\} \tag{10}$$

for n = 0,1,2,.... The next Lemma then immediately follows from the equations of (10).

Lemma 3: Let $(M(n)/M(n)/1)_i$ be a birth-death queueing model with arrival rates $\underline{\lambda}_i = (\lambda_{in})_{n=0}^\infty$ and service rates $\underline{\mu}_i = (\mu_{in})_{n=1}^\infty$, for i = 1 and 2. If $\lambda_1 \geq \lambda_2$ and $\underline{\mu}_1 \leq \underline{\mu}_2$, then

$$F_n(\underline{\lambda}_1, \underline{\mu}_1) \leq F_n(\underline{\lambda}_2, \underline{\mu}_2), \quad n=0,1,\ldots.$$

Combining Theorem 1 and Lemma 3, one obtains the following.

Theorem 4: The number of parts in the types of FMS under consideration, under any of the above three input control policies, is stochastically minimized by balancing the workload on all machines.

Note that the statement of Theorem 4 is heuristic for Policy III, since in this case, it is the approximation, which was given by equation (7), that is used to obtain the result.

5. CONCLUSIONS

In this paper, we have established the superiority of balancing workloads in flexible manufacturing systems utilizing no pooling, both to stochastically minimize the number of parts in the system as well as to maximize the expected production. Three policies to control the input of parts into the system are considered. There is a Poisson arrival process to a finite or infinite common input buffer area, exponentially distributed service times, Markovian part transfers from machine to machine, and an ample buffer at each machine.

The optimality of balanced workloads is established with the constraints of single machine work centers and $\sum_{i=1}^{M} x_i$ is a constant. When these conditions are relaxed, the balanced load need not be optimal. On the other hand, for similar systems, but with unbalanced pools, expected production is maxi- by an unbalanced load, again with the constraint that $\sum_{i=1}^{M} x_i$ is a constant. (See Stecke (1981) and Stecke and Solberg (1985).)

Yao (1984a,b) and Yao and Kim (1984) have obtained similar balancing results for cases where the above two contraints are relaxed. For example,

they show that balancing maximizes expected production when there is pooling but with the same number of machines in each group. This is also observed in Stecke (1981) and Stecke and Solberg (1985).

Empirical studies using both simulation and approximations are currently underway to extend these results to more general cases.

Additional work is required to implement a balancing workload objective over time. The FMS balancing problem is different than that of a flow shop or job shop in several ways. A flow shop is balanced once, during its design. At the other extreme, it's very difficult to balance a job shop. Usually the work is given, to result in one particular bottleneck machine type, which changes over time.

An FMS can produce in unit batch sizes. There are planning decisions to be made that impact balancing (or unbalancing). These include: selecting the part types to be produced next; determining the ratios at which these part types are to be produced; allocating pallets and fixtures among the part types; determining the minimum number of pallets required in the system; loading tools and assigning operations to machines; determining the appropriate input sequence into the FMS; and finally, the actual scheduling of parts through the system. More information concerning what these problems are for an FMS can be found in Stecke (1983, 1985).

Some models that can be used to address these problems are overviewed in Suri (1984). An efficient algorithm to determine the most balanced allocations for the types of FMSs considered in this paper is provided in Berrada and Stecke (1985). The algorithm also applies to systems of pooled machines, when each pool contains the same number of machines. Extensions to unbalanced groupings (hence unbalanced loadings) are underway. In any case, further

research is required to discover methods to address all of the above-mentioned problems in order to implement either balancing (or an unbalancing) operational objective.

ACKNOWLEDGMENT

Kathryn E. Stecke's work was supported in part by NSF Grant No. ECS 8406407.

REFERENCES

[1] Berrada, M., and K. E. Stecke, (1985), "A Branch and Bound Approach for Machine Loading in Flexible Manufacturing Systems," Management Science, forthcoming.

[2] Buzacott, J. A., (1982), "'Optimal' Operating Rules for Automated Manufacturing Systems," IEEE Transactions on Automatic Control, Vol. AC-27, pp. 80-86.

[3] Buzacott, J. A., and J. G. Shanthikumar, (1980), "Models for Understanding Flexible Manufacturing Systems," AIIE Transactions, Vol. 12, No. 4, pp. 339-350.

[4] Shanthikumar, J. G., (1982), "On the Superiority of Balanced Load in a Flexible Manufacturing System," Technical Report, Department of IE & OR, Syracuse University, New York.

[5] Shanthikumar, J. G., and R. G. Sargent, (1981), "A Hybrid Simulation/ Analytic Model of a Computerized Manufacturing System," Proceedings of the 9th IFORS Conference, Hamburg, W. Germany, pp. 901-915.

[6] Solberg, J. J., (1977), "A Mathematical Model of Computerized Manufacturing Systems," Proceedings of the International Conference on Production Research, Tokyo, Japan.

[7] Stecke, K. E., (1981), Production Planning Problems for Flexible Manufacturing Systems, Ph.D. dissertation, School of Industrial Eng., Purdue University, W. Lafayette, Indiana.

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

[9] Stecke, K. E., (1985), "A Hierarchical Approach to Solving Machine Grouping and Loading Problems of Flexible Manufacturing Systems," European Journal of Operational Research, to appear.

[10] Stecke, K. E., and T. L. Morin, (1982), "Optimality of Balanced Workloads in Flexible Manufacturing Systems," Working Paper No. 289, Graduate School of Business Administration, The University of Michigan, Ann Arbor, Michigan.

[11] Stecke, K. E., and T. L. Morin, (1984), "The Optimality of Balancing Workloads in Certain Types of Flexible Manufacturing Systems," European Journal of Operational Research, to appear.

[12] Stecke, K. E., and J. J. Solberg, (1985), "The Optimality of Unbalancing Both Workloads and Machine Group Sizes in Closed Queueing Networks of Multi-Server Queues," Operations Research, to appear.

[13] Suri, R., (1983), "Robustness of Queueing Network Formulas," Journal of the Association for Computing Machinery, Vol. 30, No. 3, pp. 564-594.

[14] Suri, R., (1984), "An Overview of Evaluative Models of Flexible Manufacturing Systems," Proceedings of the First ORSA/TIMS FMS Conference, Ann Arbor, Michigan.

[15] Yao, D. D., (1984a), "Majorization and Arrangement Orderings in Open Networks of Queues," Technical Report, Department of Ind. Eng., Columbia University, New York.

[16] Yao, D. D., (1984b), "Some Properties of the Throughput Function of Closed Networks of Queues," Technical Report, Dept. of Ind. Eng., Columbia University, New York.

[17] Yao, D. D., and S. C. Kim, (1984), "Some Order Relations in Closed Networks of Queues with Multi-Server Stations," Technical Report, Dept. of Ind. Eng., Columbia University, New York.