MANAGEMENT SCIENCE
Vol. 37, No. 4, April 1991
Printed in U.S.A.

OPTIMIZING PROCESSING RATES FOR FLEXIBLE MANUFACTURING SYSTEMS*

PAUL J. SCHWEITZER AND ABRAHAM SEIDMANN
William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, New York 14627

This paper introduces the generic concept of processing rates as decision variables in Flexible Manufacturing Systems (FMS's). The objective is to determine the minimum cost processing rates given the FMS throughput target, the work-in-process level, part routes, transporter delays, and the variable capacity cost function for each machine. A nonlinear Mean Value Analysis queueing network optimization methodology is developed to control bottlenecks and queue lengths as the processing rates are varied. This methodology further provides the average and marginal unit production costs along with necessary and sufficient feasibility conditions for the FMS throughput targets. Industrial sample data is then used to illustrate the solution of the optimal tool speed problem in a metal-cutting FMS. Considerable cost savings are demonstrated using the proposed methodology in contrast with the conventional one-machine optimization models. Several economic insights regarding the issues of capacity allocation for FMS's, and a generalization of the square root capacity allocation rule for closed networks of queues, are also presented.
(FLEXIBLE MANUFACTURING SYSTEMS; PROCESSING RATES; TOOL SPEEDS)

1. Introduction

Most production planning and control schemes for Flexible Manufacturing Systems (FMS's) treat the processing rates as given (Stecke and Suri 1986, Vollmann et al. 1988), while in practice these are significant decision variables. In order to achieve the desired production goals with minimal operating costs, managers regularly vary the aggregate plant capacity by modifying the input levels of direct labor hours and other productive resources (Bitran and Tirupati 1987, Shanthikumar and Yao 1987); similarly, process planners are instructed to concurrently vary other key manufacturing parameters (i.e., tool feed rate or cutting speeds) to reach the same objectives (Halevi 1980). The latter case—which seems to be the prevalent option for managing capacity changes in unmanned/automated facilities—is based on some well-known empirical relationships between tool or die costs and the processing rate (Drozda and Wick 1983, Taylor 1907). For example, the drilling operation analyzed in Conrad and McClamrock (1987, equation (3.4.1)) has

$$\text{mean tool life} = \text{constant}\left(\frac{1}{\text{feed rate}}\right)^{5.15},$$

so a 10% change in feed (i.e., processing) rate causes a 50% change in tool cost. This property, together with the empirical observation that tool costs can comprise 20–30% of the total operating costs of FMS (Ayres 1988, Cumings 1986), suggests that significant economies may be possible via judicious selection of processing rates.

When production targets are within the limits of the plant capacity and the unit revenue is constant, the minimum production cost criteria also leads to profit maximization. The problem of economic selection of optimum cutting conditions and of the relevant choice of cost parameters has been recently investigated by Boucher (1987), Chang et al. (1982), and Trappey et al. (1987), among others. Their models considered a single part type

* Accepted by Stephen C. Graves; received November 29, 1988. This paper has been with the authors 3 months for 2 revisions.

0025-1909/91/3704/0454$01.25

produced on a single machine. Extensions to the two-stage and to the multistage (flowshop) cases were presented by Hitomi (1971, 1979), Koulamas et al. (1987), McCartney and Hinds (1982), and Yellowley (1983). However, in many of the commercial process planning systems, the machining conditions are still determined for one machine at a time (Chang and Wysk 1985, Tipnis et al. 1987). The dynamic interaction effects, due to random part flows between machines, machine starvation, queues, WIP buildup, and bottlenecks are ignored.

Optimizing the process rates in FMS's requires a queueing network model to determine global changes in bottlenecks and queue lengths as the processing rates are altered. Our methodology extends the Mean-Value-Analysis (MVA) framework of FMS introduced by Hildebrant (1980) and by Suri and Hildebrandt (1984), which was shown to predict throughputs, queue lengths, waiting times, and machine utilizations with satisfactory accuracy for planning purposes (Seidmann et al. 1987a, b, Shalev-Oren et al. 1985, and Snowdon and Ammons 1988). Successful industrial application of similar models are reported by Brown (1988), Johnson (1989), Solberg (1978), and Suri and Hildebrant (1984).

The optimization problem over processing rates then becomes an optimization problem superimposed on top of the MVA queueing model. This optimization problem assumes that steady-state throughput objectives must be achieved (otherwise the minimum-cost solution is to produce nothing). The decision variables in the optimization model are the processing rates (or, equivalently, processing times) of each part at each machine. It will be seen that the optimization problem is complex due to the *interference effects between machines*: lowering speeds on nonbottleneck machines will shift queues and alter the location of bottlenecks, possibly forcing us to increase speeds elsewhere in order to still achieve the same throughputs. The role of the network-of-queues model is to permit a systematic analysis of these tradeoffs, which was absent from the FMS analyses cited above.

Optimization of queueing networks is frequently encountered in the design and management of various manufacturing systems (Buzacott and Yao 1986, Dallery and Frein 1986, Vinod and Sabbagh 1986, Vinod and Solberg 1985, and Wein 1989). The allocation of a given productive capacity is discussed by Akyildiz (1988), Bitran and Tirupati (1987), Pittipati and Teele (1987), and Shanthikumar and Yao (1987).

This paper analyzes the case of a homogeneous family of parts with general parts routes. It minimizes the expected cost per unit time subject to throughput and process time constraints. This is equivalent *to minimizing the cost per part, subject to throughput and processing time constraints*. The next section presents the mathematical model formulation. The algorithmic solution methodology is referenced in §3. The economic marginal returns for incremental throughput changes are given in §4. Several numerical examples are discussed in §5, and §6 concludes our paper.

2. Mathematical Model Formulation

2.1. *Introduction*

This section formulates the problem of optimizing processing rates for the case of one part type. §2.2 gives the assumptions and notation, §2.3 incorporates the modelling aspects of the material handling system, §2.4 gives the mean-value analysis and optimization problem, and §2.5 gives a transformed version more convenient for optimization purposes.

2.2. *Assumptions and Notations*

The model inputs are K, M, v_i, TH, $s_i^{\mp}$, D_i, and $g_i(\cdot)$.

$-K \geq 1$ pallets (K = WIP inventory level).

–M reliable machines labelled $i = 1, 2, \ldots, M$. (All machines are FCFS. Each machine can process only one part type at a time and is never blocked.)

–Machine 1 is the load/unload station (L/UL).

–v_i = visit ratio at machine i, $i = 1, 2, \ldots, M$; = number of times a part visits machine i before being completed and leaving at the L/UL station. It is assumed that $v_1 = 1$ and that every visit ratio is strictly positive (otherwise delete this machine).

–TH = desired throughput (parts/hour going through L/UL station).

The unknown quantities of interest are

–λ_i = part throughput at machine i (production flow rate in parts/hour).

–N_i = mean number of parts at machine i (either in process or waiting for the machine).

–W_i = mean sojourn time at machine i (either in process or waiting for the machine).

–s_i = mean processing time at machine i (= time in hours for one operation).

The $\{s_i\}$ are our decision variables and must satisfy

$$0 < s_i^- \le s_i \le s_i^+, \qquad i = 1, 2, \ldots, M, \tag{2-1}$$

where the limits $s_i^{\mp}$ are set by technological or management considerations. The process time includes both part setup time and machine operation time.

We also assume knowledge of the variable capacity cost functions:

–$g_i(s_i)$ = cost of doing *one* operation on machine i, if the process time per operation is s_i.

Following earlier studies (Hax and Candea 1984, Hitomi 1979, and Johnson and Montgomery 1974) it is assumed that

$$\frac{dg_i(s_i)}{ds_i} < 0, \qquad \frac{d^2 g_i(s_i)}{ds_i^2} > 0, \qquad s_i^- \le s_i \le s_i^+. \tag{2-2}$$

This common assumption implies that increasing costs are incurred for increasing processing rates (decreasing processing times), for example, through the incremental use of overtime labor. In the case of automated machining operations, the cost function may involve tool replacement or regrinding (due to wear or breakage), labor for tool mounting, machine parts deterioration, maintenance, and any variable supervision cost. Our model assumes $Dg_i(s_i) < 0$, which is typical for a capital-intense environment where tool costs are of major concern (Primrose and Leonard 1986). Labor costs for supervision, parts remounting, etc., depend primarily on the *fixed* FMS throughput, not on the variable processing rates at the individual machines. A more detailed discussion of the use of such cost functions in production management context is given by Hax and Candea (1984). The cost function $g_i(s_i)$ implicitly includes lower-level decisions about the single operation at machine i. For example, it may involve a joint optimization of several suboperations over tool choices, turning rate (speed), and tool feed rate, subject to a constraint that the total machine time is s_i. Also, this formulation can handle the case of parallel, identical or heterogeneous machines with given visit ratios at each one.

2.3. *Modelling the Material Handling System*

The model includes a *known* mean transport delay D_i (≥ 0) associated with bringing a part into and out of work center i, $1 \le i \le M$. These transport delays may be calculated from an auxiliary queueing model which predicts the delay $D(\bar{\lambda})$ at a transporter as a function of the imposed load $\bar{\lambda}$ on the transporter. The value of $\bar{\lambda}$ is shown by summing (2-9) over all machines, i, which use this transporter. The values of $\bar{\lambda}$ and, therefore, D_i are *unchanged* when optimizing the processing rates subject to the FMS throughput constraint. They must be checked to neither exceed the capacity of the transporter, nor cause significant blocking at the output port of the work stations

2.4. *Mean-Value Analysis and Optimization Problem*

The optimization problem may be formulated as

$$\text{COSTMIN} = \text{Min} \sum_{i=1}^{M} \lambda_i g_i(s_i) \tag{2-3}$$

subject to

$$0 < s_i^- \le s_i \le s_i^+, \qquad i = 1, 2, \ldots, M, \tag{2-4}$$

$$\lambda_i = Kv_i / \sum_{j=1}^{M} v_j(W_j + D_j), \qquad i = 1, 2, \ldots, M, \tag{2-5}$$

$$W_i = s_i + [N_i(K-1)/K]s_i, \qquad i = 1, 2, \ldots, M, \tag{2-6}$$

$$N_i = \lambda_i W_i, \qquad i = 1, 2, \ldots, M, \tag{2-7}$$

$$\lambda_1 = TH. \tag{2-8}$$

These may be interpreted as follows: The objective function COSTMIN is the total operating cost per hour, with the ith summand showing λ_i parts/hour visiting machine i and costing $g_i(s_i)$ per visit. Constraint (2-4) gives the minimum and maximum allowed processing time on machine i. Equations (2-5)–(2-7) are the MVA equations (Schweitzer 1979, Shalev-Oren et al. 1985, and Suri and Hildebrant 1984) for a closed network-of-queues with M servers, population $K \ge 1$ and one customer class. Equation (2-5) expresses λ_i as the visit ratio v_i times the throughput λ_1 at the L/UL. The latter is, by Little's law, the population K divided by the mean sojourn time for a part in the system. Note each visit to machine j involves a mean time $(W_j + D_j)$ for both sojourn and transport. Equation (2-6) expresses the mean sojourn time at machine i as the sum of the mean processing time s_i plus the mean queueing delay. The latter is estimated by the ergodic mean work backlog $N_i s_i$ multiplied by a $(K-1)/K$ correction to preclude queueing for oneself (Schweitzer et al. 1986). Equation (2-7) is Little's law. Equation (2-8) is the constraint that throughput at the L/UL station is TH.

Some simplifications of (2-3)–(2-8) are possible. First rewrite (2-5) and (2-8) as

$$\lambda_i = v_i TH, \qquad i = 1, 2, \ldots, M. \tag{2-9}$$

Next, eliminate N_i via (2-7) and (2-9), permitting us to rewrite (2-6) as

$$W_i = \frac{s_i}{1 - (K-1)v_i s_i TH/K}, \qquad i = 1, 2, \ldots, M. \tag{2-10}$$

Also, (2-9) and (2-10) permit us to eliminate W_i and λ_i and rewrite (2-5) as

$$\frac{TH}{K} \sum_{i=1}^{M} \frac{v_i s_i}{1 - (K-1)v_i s_i TH/K} = y_0, \tag{2-11}$$

where

$$y_0 \equiv 1 - \sum_{i=1}^{M} \frac{TH v_i D_i}{K}$$

is known. The problem is feasible only if $y_0 > 0$. Otherwise the transport delays are too large to permit achieving the desired throughput TH. Note that y_0 is the fraction of the parts at the work centers (not at the transporter).

The utilization $v_i s_i TH$ of work center i must be ≤ 1. This permits us to *replace* s_i^+ by

$$s_i^+ = \min [s_i^+, 1/(v_i TH)], \qquad i = 1, 2, \ldots, M, \tag{2-12}$$

which in turn ensures that the inequality

$$(K-1)v_i s_i TH/K < 1, \qquad i = 1, 2, \ldots, M, \tag{2-13}$$

needed for (2-11), will be met if $s_i \le s_i^+$.

In summary, the processing rates $\{s_i\}$ are chosen to minimize the total cost rate

$$\text{COSTMIN} = \sum_{i=1}^{M} THv_i g_i(s_i)$$

subject to the bounding constraints (2-4) (with revised definition of s_i^+) and the throughput constraint (2-11).

THEOREM 1. *The problem is feasible if and only if the following three conditions hold*:

$$0 < s_i^- \le s_i^+, \qquad i = 1, 2, \ldots, M \ (\textit{using revised } s_i^+), \tag{2-14a}$$

$$\sum_{i=1}^{M} \frac{THv_i D_i}{K} < 1 \qquad (\textit{so } y_0 > 0), \tag{2-14b}$$

$$\frac{TH}{K}\sum_{i=1}^{M}\frac{v_i s_i^-}{1-(K-1)v_i s_i^- TH/K} \le y_0$$

$$\le \frac{TH}{K}\sum_{i=1}^{M}\frac{v_i s_i^+}{1-(K-1)v_i s_i^+ TH/K} \qquad (\textit{using revised } s_i^+). \tag{2-14c}$$

See Schweitzer and Seidmann (1988) for the proof.

Also note from Schweitzer and Seidmann (1988) that the range of feasible FMS throughputs is $TH^- \le TH \le TH^+$, where TH^+ is given *uniquely* by

$$1 = \frac{TH^+}{K}\sum_{i=1}^{M} v_i\left\{D_i + \frac{s_i^-}{1-(K-1)v_i s_i^- TH^+/K}\right\} \tag{2-15}$$

and

$$\frac{(K-1)v_i s_i^- TH^+}{K} < 1, \qquad i = 1, 2, \ldots, M, \tag{2-16}$$

and TH^- is given *uniquely* by

$$1 = \frac{TH^-}{K}\sum_{i=1}^{M} v_i\left\{D_i + \frac{s_i^+}{1-(K-1)v_i s_i^+ TH^-/K}\right\} \tag{2-17}$$

and

$$\frac{(K-1)v_i s_i^+ TH^-}{K} < 1, \qquad i = 1, 2, \ldots, M. \tag{2-18}$$

The uniqueness of these throughput bounds arises from the strict monotonicity of the right-hand sides of (2-15) and (2-17).

Furthermore, if the number of pallets, K, can be varied, a given choice of K is feasible for the throughput goal TH if and only if $K^- \le K \le K^+$, where

$$K^- = \min\{K \text{ integer} \mid TH^+ \ge TH\}, \tag{2-19}$$

$$K^+ = \max\{K \text{ integer} \mid TH^- \ge TH\}. \tag{2-20}$$

These properties hold because throughput is *monotone* in K.

2.5. *Transformed Formulation*

This subsection gives a monotone change of variables from s_i to x_i which simplifies the original problem formulation.

The monotone change of variables is

$$x_i \equiv \frac{v_i s_i TH/K}{1 - v_i s_i TH(K-1)/K}, \qquad i = 1, 2, \ldots, M. \tag{2-21}$$

Define

$$f_i(x_i) \equiv v_i TH g_i\left(s_i = \frac{x_i K/(v_i TH)}{1 + x_i(K-1)}\right)$$

and

$$x_i^+ \equiv \frac{v_i s_i^+ TH/K}{1 - v_i s_i^+ TH/K}, \qquad x_i^- \equiv \frac{v_i s_i^- TH/K}{1 - v_i s_i^- TH/K}, \qquad i = 1, 2, \ldots, M \tag{2-22}$$

(using revised s_i^+). The properties

$$\frac{df_i(x_i)}{dx_i} < 0, \qquad \frac{d^2 f_i(x_i)}{dx_i^2} > 0, \qquad i = 1, 2, \ldots, M, \tag{2-23}$$

are inherited from (2-2), and the bounds $x_i^- \leq x_i \leq x_i^+$, $1 \leq i \leq M$, follow from (2-4) and the monotonicity of (2-21).

The revised optimization problem is then

$$\text{COSTMIN} = \min \sum_{i=1}^{M} f_i(x_i) \tag{2-24}$$

subject to

$$x_i^- \leq x_i \leq x_i^+, \qquad i = 1, 2, \ldots, M, \tag{2-25}$$

$$\sum_{i=1}^{M} x_i = y_0 \tag{2-26}$$

with property (2-23) and the feasibility conditions from Theorem 1:

$$x_i^- \leq x_i^+, \qquad i = 1, 2, \ldots, M, \tag{2-27a}$$

$$y_0 > 0 \qquad (\text{cf. } (2\text{-}14b)), \tag{2-27b}$$

$$\sum_{i=1}^{M} x_i^- \leq y_0 \leq \sum_{i=1}^{M} x_i^+. \tag{2-27c}$$

3. Algorithmic Solution

The model formulated in (2-24)–(2-26) is a convex resource-allocation problem, with bounded variables. If feasible, this problem possesses a unique optimal solution due to the strict convexity of the objective function. Earlier published treatments of such resource-allocation problems include (Bitran and Hax 1981, Ibaraki and Katoh 1988, Luss and Gupta 1975, and Zipkin 1980). Since M is small, we found it convenient to develop an alternate optimization algorithm. The coded algorithm required less than one minute of an IBM/PC for each one of the problems analyzed in the next section (Schweitzer and Seidmann 1988).

4. Marginal Capacity Costs

4.1. *Scope*

In §4.2 the marginal returns are given for the optimal processing rates. Upper and lower bounds on the rate of change of the unit production cost per unit change in the system throughput constraint are then given in §4.3. Additional economic insights regarding the optimal capacity allocation issues are discussed in §4.4.

4.2. *The Marginal System Cost Rate*

The marginal system cost is defined by:

$$\beta \equiv \partial(\text{COSTMIN})/\partial(TH). \tag{4-1}$$

The value of β is not available directly because (2-8) was used directly to eliminate all throughputs λ_i via (2-9). However, Theorem 2 presents an alternative system model which also provides an appealing structural insight.

THEOREM 2. *For a desired throughput level TH, the margin system cost rate is given by:*

$$\beta = \frac{\text{COSTMIN}}{TH} - \frac{b*}{TH}\left[1 + (K-1)\sum_{i=1}^{M}(x_i)^2\right], \tag{4-2}$$

where the optimal values of COSTMIN *and x_i are used. Here $b*$ denotes the common value of $df_i(x_i)/dx_i$ for all optimal x_i values which are interior to their feasible range $[x_i^-, x_i^+]$. The transport delays are assumed to be load independent ($\partial D_i/\partial TH = 0$).*

(See Schweitzer and Seidmann 1988 for a detailed proof.)

Note that since $b^* < 0$, $\beta > 0$ as expected: increasing the throughput *TH* increases the total cost rate. The β value equals the optimal average production cost per unit (COSTMIN/*TH*) plus an added term which lies between one and *K* multiples of b^*/TH. Its value provides a *lower bound* on the cost increment when an increase in the throughput goal is considered. This holds because β is a sharply increasing function of *TH* due to the accelerated tool burnout.

4.3. *Marginal Unit Production Costs*

The (minimal) *cost per part* (unit cost), when the throughput is *TH*, is given by

$$\tilde{c} = \frac{\text{COSTMIN}}{TH}. \tag{4-3}$$

The change in the unit cost $\tilde{c}$ per unit change in throughput is

$$\frac{\partial \tilde{c}}{\partial(TH)} = \frac{1}{TH}\frac{\partial(\text{COSTMIN})}{\partial(TH)} - \frac{\text{COSTMIN}}{(TH)^2} = \frac{\beta}{TH} - \frac{\text{COSTMIN}}{(TH)^2}.$$

Using (4-2), this is

$$\frac{\partial \tilde{c}}{\partial(TH)} = \frac{-b^*}{(TH)^2}\left[1 + (K-1)\sum_{i=1}^{M}(x_i)^2\right]. \tag{4-4}$$

This reveals the central role of b^* in determining the sensitivity of unit costs to throughput.

4.4. *Economic Insights*

Further understanding of the optimal capacity allocation policy can be provided through the analysis of a special case in which (2-4) is replaced by

$$0 \le s_i \tag{4-5}$$

and using the following cost functions:

$$g_i(s_i) = c_i/s_i, \qquad i = 1, 2, \ldots, M \tag{4-6}$$

(i.e., cost per operation is linear in capacity $1/s_i$).

The optimal processing times in this case can be *explicitly* derived by optimizing the Lagrangian function for (2-3) with the throughput constraint derived from (2-11), and assuming that the transporter delays (D_i's) are independent of TH.

In this case the optimal processing rates are given by

$$\frac{1}{s_i} = THv_i\left(1 + \frac{\sum_{j=1}^{M} v_j\sqrt{c_j}}{v_i\sqrt{c_i}Ky_0} - \frac{1}{K}\right). \tag{4-7}$$

Our result (4-7) generalizes the classical square root capacity allocation rule (Kleinrock 1964) for *closed* queueing network models of FMS's with transporter delays. It also differs from Akyildiz (1988) and Pattipati and Teele (1987) as our original objective function (2-3) incorporates the cost effects of distinct part visits to machines.

The relationship in (4-7) indicates that the optimal processing rate at machine i equals the minimal capacity v_iTH—needed at this machine in order to meet the throughput goal—plus an added nonnegative increment; the percentage increment in this machine capacity is inversely proportional to the product of the visit ratio times the square root of the marginal capacity cost rate parameter. Note, in particular, that the distribution of the incremental (or excess) capacity across the various machines is inversely proportional to the relative fraction of $v_i\sqrt{c_i}$'s, and that it diminishes with the increase in K. The *bottleneck* machine (i.e., highest utilization $= s_iTHv_i$) is the one associated with the *highest* $v_i\sqrt{c_i}$. In general, (4-7) indicates that it is usually *not* optimal to set equal utilization at all machines, or equal waiting times at all machines, or to use the process times to compensate for the transporters delays.

The waiting times are given by

$$W_i = \frac{y_0K}{TH\sum_{j=1}^{M} v_j\sqrt{c_j}}\sqrt{c_i} = \text{constant}\sqrt{c_i}, \qquad i = 1, 2, \ldots, M. \tag{4-8}$$

Hence the mean waiting time for machine i is proportional to the values of $\sqrt{c_i}$. Machines with lower marginal capacity costs will be associated with shorter WIP queues. The optimal waiting time will be *identical* at all machines (regardless of their visit ratios and transporter delays) if their marginal capacity cost rates are identical.

These insights manifest the relative importance of *both* $\sqrt{c_i}$ and $v_i\sqrt{c_i}$—depending upon the operational issue in question.

Finally, the marginal production cost rate is given by:

$$\beta = TH\left[2\frac{K-1}{K}\sum_{i=1}^{M} c_iv_i^2 + \frac{(\sum_{i=1}^{M} v_i\sqrt{c_i})^2 K(y_0+1)}{(Ky_0)^2}\right]. \tag{4-9}$$

5. Numerical Applications

We use here a typical data set for a metal-cutting FMS with six machines and one L/UL station ($M = 7$). This facility is designed to operate with 21 ($=K$) pallets. Table 1 presents the process requirements for all machines. The nominal process times—depicted for this application by Table 1—were computed from a commonly used industrial stand-alone minimum cost machining model for iron base alloy (H-46) parts using special carbide tools (see Schweitzer and Seidmann 1988 for further details).

The following investigations focus on *tool cost reduction* because labor costs are fixed once the throughput target is specified. Table 2 presents the parameters of the tool cost functions $g_i(s_i) = a_is_i^{-b_i}$ and the feasible range of the processing times (s_i^-, s_i^+) for all

TABLE 1
Process Requirements

Machine (i)	Nominal Process Time (s_i) (min)	Transportation Time (D_i) (min)	Visit Ratio (v_i)
1	1.00	10.0	1.0
2	12.17	7.0	1.0
3	16.24	8.0	1.0
4	14.40	5.0	1.4
5	18.01	5.0	0.6
6	6.51	3.0	0.7
7	6.51	3.0	0.3

TABLE 2
Range of Feasible Process Times and Tool Cost Functions Parameters

Machine	Process Times Range		Tool Cost Function	
(i)	s_i^- (min)	s_i^+ (min)	Coefficient (a_i)	Exponent (b_i)
1	1.00	1.00	—	—
2	9.11	15.19	3465	3.166
3	10.86	21.00	4415	2.410
4	4.60	15.35	9445	4.190
5	4.60	20.21	9445	4.190
6	3.91	11.20	6434	3.140
7	3.91	11.20	6434	3.140

TABLE 3
FMS Performance Measures Before and After Process Times Optimization

Machine (i)	Throughput (λ_i) (products/min)	Process Times (s_i) (min)		Sojourn Time (w_i) (min)		Utilization (%)	
		Nominal	Optimal	Nominal	Optimal	Nominal	Optimal
1	0.0483	1.00	1.00	1.048	1.048	4.8	4.8
2	0.0483	12.17	15.19	27.623	50.422	58.7	73.4
3	0.0483	16.24	19.78	64.083	219.065	78.4	95.5
4	0.0676	14.40	12.43	196.397	62.238	97.3	84.0
5	0.0290	18.01	17.81	35.786	35.040	52.2	51.6
6	0.0338	6.51	11.20	8.735	17.518	22.0	37.9
7	0.0145	6.51	11.20	7.152	13.248	9.4	16.2

Initial cost per product = $\tilde{c}$ = 24.76 (dollars/part).
Optimized cost per product = $\tilde{c}$ = 7.59 (dollars/part).
Marginal production cost = β = 17.93 (dollars/part).
Marginal unit production cost $(\partial\tilde{c}/\partial TH)$ = 214.1 (dollars − min/part^2).

machines. These ranges are a function of the tool vendors recommendations (Drozda and Wick 1983, Wagner and Barash 1971). Given these industrial processing ranges, the range of feasible throughputs is: $TH^+ = 0.0888$ parts/min to $TH^- = 0.0436$ parts/min. A queueing MVA determined that the expected throughput, using the nominal processing rates for this FMS, should be 0.0483 parts/min

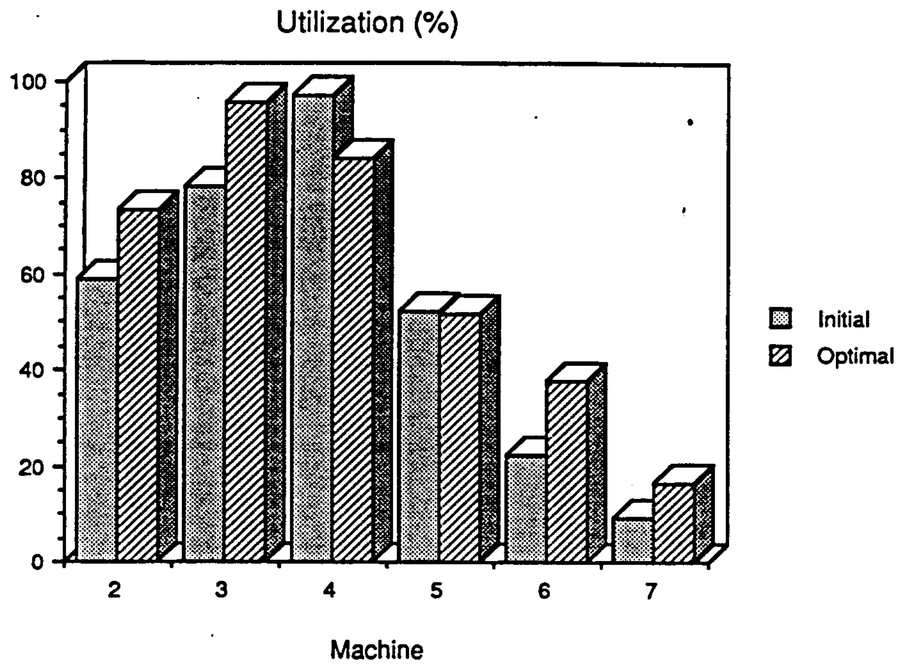

FIGURE 1. The Nominal and Optimal Machine Utilization.

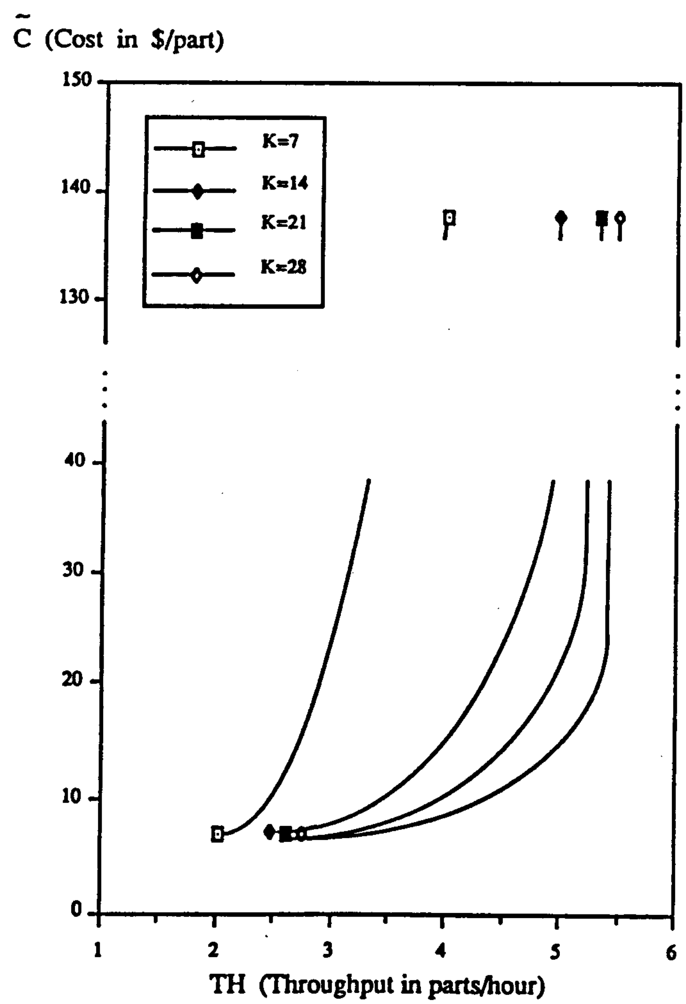

FIGURE 2. The Optimal Cost Per Part $\tilde{c}$ as a Function of the Throughout Target TH and the WIP Level K.

Assuming a throughput goal of 0.0483 parts/min, the optimization model is now used to reduce tool costs. Following the solution of (2-24)–(2-26), the tool costs are reduced from \$24.76 to \$7.59 per product. *This factor of* 3 *reduction in tool costs might translate into a* 5–15% *reduction in unit costs*. In other experiments we performed, *similar significant reductions were also achieved*.

Table 3 and Figure 1 explain the tool cost savings. The bottleneck machine 4 was *slightly* accelerated while the processing rates at all other machines (except machine 5) were significantly reduced—thereby providing for substantial gains in their tool lives. Note that machines 4 and 5, with *identical cost functions*, end up with different process times as a result of having *distinct visit ratios* (workloads). This outcome is clearly different from what the commonly used one-machine model would suggest. Figure 2 explores the sensitivity of unit cost ($\tilde{c}$) to changes in both WIP values (K) and throughput target (TH). These results suggest that $\tilde{c}$ increases very sharply with TH, and that increasing K reduces $\tilde{c}$ for fixed TH. Increasing K both enlarges the range $[TH^-, TH^+]$ of feasible throughputs and may be used to increase throughput without sacrificing tool costs.

6. Summary and Conclusions

An optimization methodology is presented to determine the desirable processing rate for each machine in an FMS. The optimization framework used is superimposed on top of an MVA queueing network model which captures the dynamic effects on bottlenecks and queue lengths as the processing rates are changed.

Analyzing the structure and performance of the FMS with optimized processing rates leads to the following observations:

(a) In the special case of *linear* cost rate functions, the optimal processing rate at each machine equals the minimal capacity needed in order to meet the throughput goal, plus an added nonnegative increment. The percentage increment in this machine capacity is inversely proportional to the product of its visit ratio times the square root of the marginal capacity cost parameter.

(b) In the above case, the bottleneck machine is the one having the highest value of the part visit ratio times the square root of the marginal capacity cost parameter. The optimal mean sojourn times at each machine are also proportional to the square root of the marginal capacity cost parameter.

(c) It is not generally optimal to seek equal utilization of all machines, or equal waiting times at all machines, or to use the process times to compensate for transporter delays.

(d) Considerable savings in tool cost (and in operating costs in general) are realized using the proposed queueing network optimization methodology compared with the conventional one-machine process planning models. The latter ignore system-wide issues such as the relative workload on each machine, the FMS throughput goal, and transporter delays. Optimization leads to slight acceleration of the processing rates at a few (economic bottleneck) machines and allows for significant processing rate reduction at many others—thereby providing for substantial gains in tool lives.

(e) Increasing the WIP level K can significantly increase the throughput range and reduce the operating costs for a given throughput goal, but at the cost of increased lead times.

(f) The marginal part production costs tends to grow sharply as the throughput target approaches the upper bound of the feasible range. Recognizing this structural property is central for making justification decisions regarding requested capacity expansions, for setting economic throughput levels, and for assessing the impact of demand uncertainty on the corporate revenues. It also suggests that there are significant economic benefits from *throughput smoothing* if the throughput targets imposed on an operating FMS are alternately high and low.

The methodology presented above expands the decision scope of operations managers as well as process planners. In the short run, distinct processing rates and part routes should be periodically updated by the operations manager in order to effectively accommodate external changes in the master schedule throughput levels, and the availability of machines, tools, and labor. In the long run, this methodology is useful for analyzing capacity allocation decisions such as adding machines to work centers, determining pallet numbers, and selection of material handling systems. In particular, we found it useful in evaluating the tradeoff between reducing variable production costs and increasing both work-in-process inventories and leadtimes.[1]

[1] The authors would like to thank Professor M. M. Barash from Purdue University, along with Professor H. Mendelson from Stanford University, for useful discussions, and Paulo Goes from the University of Rochester for his help in scientific programming. Partial support for this study was provided by the IBM Program of support for Education in the Management of Information Systems and by the Simon School Center for Manufacturing and Operations Management at the University of Rochester.

References

AKYILDIZ, I. F., "On Optimal Performance Measures of Computer Systems," Technical Report GIT-ICS-88/16, Georgia Institute of Technology, School of Information and Computer Science, Atlanta, GA, 1988.

AYRES, R. V., "Future Trends in Factory Automation," *Manufacturing Rev.*, 1, 2 (1988), 93–103.

BITRAN, G. R. AND A. C. HAX, "Disaggregation and Resource Allocation Using Convex Knapsack Problems with Bounded Variables," *Management Sci.*, 27, 4 (1981), 431–441.

——— AND D. TIRUPATI, "Trade-off Curves, Targeting and Balancing in Manufacturing Networks," Working Paper WP87-08-05, Sloan School of Management, MIT, Cambridge, MA, 1987.

BOUCHER, T. O., "The Choice of Cost Parameters in Machining Cost Models," *The Engrg. Economist*, 32, 3 (1987), 217–230.

BROWN, E., "IBM Combines Rapid Modeling Technique and Simulation to Design PCB Factory of the Future," *Industrial Engrg.*, 20, 6 (1988), 23–26.

BUZACOTT, J. A. AND D. D. YAO, "Flexible Manufacturing Systems: Review of Analytical Models," *Management Sci.*, 32, 7 (1986), 890–905.

CHANG, T., R. A. WYSK, R. P. DAVIS AND B. CHOI, "Milling Parameter Optimization Through a Discrete Variable Transformation," *Internat. J. Production Res.* 20, 4 (1982), 507–516.

——— AND ———, *An Introduction to Automated Process Planning Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1985.

CONRAD, C. J. AND N. H. MCCLAMROCK, "The Drilling Problem: A Stochastic Modeling and Control Example in Manufacturing," *IEEE Trans. Automat. Control*, AC-32, 11 (1987), 947–958.

CUMINGS, S., "Developing Integrated Tooling Systems: A Case Study at Garret Turbine Engine Company," in *Proc. Fall Industrial Engineering Conf.*, Boston, MA, 1986.

DALLERY, Y. AND Y. FREIN, "An Efficient Method to Determine the Optimal Configuration of a Flexible Manufacturing System," in *Proc. Second* ORSA/TIMS *Conf. on Flexible Manufacturing Systems*, K. E. Stecke and R. Suri (Eds.), Elsevier, Amsterdam, 1986.

DROZDA, T. J. AND C. WICK (EDS.), *Tool and Manufacturing Engineers Handbook*, Vol. I, Society of Manufacturing Engineers, Dearborn, MI, 1983.

HALEVI, G., *The Role of Computers in Manufacturing*, Wiley, Nw York, 1980.

HAX A. C. AND D. CANDEA, *Production and Inventory Management*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

HILDEBRANT, R. R., "Scheduling Flexible Machinery Systems Using Mean Value Analysis," in *Proc. IEEE Conf. on Decision and Control*, Albuquerque, NM, 1980, 701–706.

HITOMI, K., "Optimization of Multistage Machining System: Analysis of Optimal Machining Conditions for the Flow-Type Machining System," *ASME, J. Engrg. Industry*, 73, 2 (1971), 598–606.

———, *Manufacturing Systems Engineering*, Taylor & Francis, London, 1979.

IBARAKI, T. AND N. KATOH, *Resource Allocation Problems: Algorithmic Approaches*, MIT Press, Cambridge, MA, 1988.

JOHNSON, E., "Manufacturing System Analysis Using Queueing Network Software," Working Paper (MS 28AQ), Hewlett-Packard Corporation, Palo Alto, CA, 1989.

JOHNSON, L. A. AND D. C. MONTGOMERY, *Operations Research in Production Planning, Scheduling and Inventory Control*, Wiley, New York, 1974.

KLEINROCK, L., *Communication Nets: Stochastic Message Flow and Delay*, McGraw-Hill, New York, 1964.

KOULAMAS, C. P., B. K. LAMBERT AND M. L. SMITH, "Optimal Machining Conditions and Buffer Space Size for the Two-Stage Case," *Internat. J. Production Res.*, 25, 3 (1987), 327–336.

LUSS, H. AND S. K. GUPTA, "Allocation of Effort Resources Among Competing Activities," *Oper. Res.*, 23, 6 (1975), 360–366.

MCCARTNEY, J. AND B. K. HINDS, "Tooling Economics in Integrated Manufacturing Systems," *Internat. J. Production Res.*, 20, 4 (1982), 493–505.

PATTIPATI, K. R. AND J. L. TEELE, "Approximate Mean Value Analysis Algorithm for Queueing Networks: New Implementations and Convergence Results (Part I: Single Class Case)," Technical Report TR-87-10, Dept. of Electrical and Systems Engineering, University of Connecticut, Storrs, CT, 1987.

PRIMROSE, P. L. AND R. LEONARD, "Reappraising Cutting Tool Economies Within the Bounds of Accountancy Theory," *Internat. J. Production Res.*, 24, 2 (1986), 269–278.

SCHWEITZER, P., "Approximate Analysis of Multiclass Closed Networks of Queues," in *Proc. Internat. Conf. on Stochastic Control and Optimization*, Free University, Amsterdam, The Netherlands, April 5–6, 1979.

——— AND A. SEIDMANN, "Models for Processing Rates Optimization in Flexible Manufacturing Systems," Working Paper QM88-14 (revised), William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, NY, 1988.

———, ——— AND S. SHALEV-OREN, "The Correction Terms in Approximate Mean Value Analysis," *Oper. Res. Lett.*, 4 (1986), 197–200.

SEIDMANN, A., P. J. SCHWEITZER AND S. SHALEV-OREN "Computerized Closed Queueing Network Models of Flexible Manufacturing Systems," *Large Scale Systems*, 12 (1987a), 91–107.

———, S. SHALEV-OREN AND P. J. SCHWEITZER, "Mean Value Analysis of Interactive Computer Systems with HOL Priority and Parallel Servers," Working Paper (QM-87-17), W. E. Simon Graduate School of Business Administration, University of Rochester, Rochester, NY, 1987b.

SHALEV-OREN, S., A. SEIDMANN AND P. J. SCHWEITZER, "Analysis of Flexible Manufacturing Systems with Priority Scheduling: PMVA," *Ann. Oper. Res.* 6 (1985), 115–139.

SHANTHIKUMAR, J. G. AND D. D. YAO, "Optimal Server Allocation in Systems of Multiserver Stations," *Management Sci.*, 33 (1987), 1173–1180.

SNOWDON, J. L. AND J. C. AMMONS, (1988), "A Survey of Queueing Network Packages for the Analysis of Manufacturing Systems," *Manufacturing Rev.*, 1, 1 (1988), 14–25.

SOLBERG, J. J., "Quantitative Design Tools for Computerized Manufacturing Systems," in *Proc. Sixth North American Metalworking Research Committee*, Gainesville, FL, 1978.

STECKE, K. E. AND R. SURI (EDS.), *Flexible Manufacturing Systems: Operations Research Models and Applications*, Elsevier, New York, 1986.

SURI, R. AND R. R. HILDEBRANT, "Modeling Flexible Manufacturing Systems Using Mean-Value Analysis," *J. Manufacturing Systems*, 1, 3 (1984), 27–38.

TAYLOR, F. W., "On the Art of Cutting Metals," *Trans. Amer. Soc. Mechanical Engineers*, 28 (1907), 31–350.

TIPNIS, V. A., S. A. VOGEL, AND C. E. LAMB, "Computer-Aided Process Planning Systems for Aircraft Engine Rotating Parts," in *Capabilities of Group Technology*, N. L. Hyer (Ed.), SME Publication, Dearborn, MI, 1987, 152–162.

TRAPPEY, J. F. C., C. R. LIU, AND T. C. CHANG, "Fuzzy Nonlinear Programming: Theory and Applications in Manufacturing," Research Report TR-ERC 87-7, School of Engineering, Purdue University, 1987.

VINOD, B. AND M. SABBAGH, "Optimal Performance Analysis of Manufacturing Systems Subject to Tool Availability," *European J. Oper. Res.*, 24, 6 (1986), 398–409.

——— AND J. J. SOLBERG, "The Optimal Design of Flexible Manufacturing Systems," *Internat. J. Production Res.*, 23, 6 (1985), 1141–1151.

VOLLMANN, T. E., W. L. BERRY AND D. C. WHYBARK, *Manufacturing Planning and Control Systems*, (Second Ed.), Irwin, Homewood, IL, 1988.

WAGNER, J. G. AND M. M. BARASH, "Study of the Distribution of the Life of HSS Tools," ASME, *J. Engrg. for Industry*, 93, 11 (1971), 1044–1050.

WEIN, L. M., "Capacity Allocation in Generalized Jackson Networks," *Oper. Res. Lett.*, 8, 3 (1989), 143–146.

YELLOWLEY, I., "A Fundamental Examination of the Economies of the Two-Pass Turning Operation," *Internat. J. Production Res.*, 21, 5 (1983), 617–626.

ZIPKIN, P. H., "Simple Ranking Methods for Allocation of One Resource," *Management Sci.*, 26, 1 (1980), 34–43.