

Performance Management in Flexible Manufacturing Systems

PAUL J. SCHWEITZER

William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, NY 14627

ABRAHAM SEIDMANN

William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, NY 14627

PAULO B. GOES

Operations and Information Management, University of Connecticut, Storrs, CT 06269

Abstract. This article treats several performance management decision problems in flexible manufacturing systems (FMSs). This work differs from a number of other studies in that we allow the processing rates at the machines to be varied, and the system has to meet a given throughput goal per unit time. The managerial decision options modeled here include part routing and allocation of tasks to machines, work-in-progress (WIP) levels, capacity expansions, tool-type selection, the setting of throughput goals, and multiperiod production planning. We discuss and explain the insights and implications, partly nonintuitive, gained from our investigations. Finally, extensive numerical evaluations are included to illustrate the economic and performance impact of the various performance management alternatives. These results demonstrate that substantial economic benefits can be achieved by careful tuning of the FMS operational parameters.

Key Words: performance management decisions, variable processing rates, economic impacts, queueing networks, capacity management

1. Introduction

Studies by Hitomi (1971, 1979), McCartney and Hinds (1982), Schweitzer and Seidmann (1988, 1989, 1991), and Watanabe and Fujii (1988) have provided frameworks for optimizing processing rates in production systems operating subject to throughput constraints. These studies were motivated by the empirical observation that tool costs comprise 20%–30% of the total operating costs of many flexible manufacturing systems (FMSs) (Ayres 1988; Cummings 1986). Significant cost savings were demonstrated for repetitive metal-cutting operations by a careful choice of the workcenters' operating parameters.

This article further extends some of these earlier results by analytical and numerical investigation of several key decision options in FMS capacity management. Numerical investigation seems necessary for developing insights into the dynamic system behavior. No analytical results are available for many of the design and operating aspects of these queueing models; the number of decision variables is large (tool feed rates and surface speeds, tool choices, FMS cell configuration, parts routing, pallet inventories, production scheduling, etc.); and the systems are tightly coupled, with changes in one set of task parameters at one workcenter strongly affecting behavior at others (floating bottleneck effects). For example, using a *system model* that takes interactions into account, Schweitzer and Seidmann

(1991) show that the current industrial practice of using single-machine models to determine economic processing rates leads to inferior decisions and significantly degraded performance.

This article begins with qualitative (section 2) and quantitative (section 3) descriptions of the FMS model. Section 4 provides the *base case* and supplies cost and performance descriptions, both before and after optimization of the processing rates.

The main body of this article consists of several analyses and case studies in which we vary the parts routes (section 5), tool choice (section 6), capacity (section 7), number of pallets (section 8), throughput goal (section 9), and production schedule (section 10). In each section we describe the decision model, performance criteria, and a detailed case experiment. We then assess the results and suggest management implications for FMS performance management. Section 11 concludes the article with our overall assessment of these results and their implications for performance tuning.

2. The FMS model

The FMS operates with $M \geq 1$ workcenters labeled $i = 1, 2, \dots, M$. Not all workcenters do machining; some do inspection, washing, etc. Each workcenter either has one server with a first-come first-serve (FCFS) service discipline (with unlimited queueing space) or is an ample server (AS) (enough parallel machines that a queue never develops). The ample-server case is useful for modeling known delays or modeling transporters such as conveyor belts. We label $i = 1$ as the load/unload station (L/UL).

Workcenter i ($1 \leq i \leq M$) can perform $n(i) \geq 1$ distinct types of operations labeled as $j = 1, 2, \dots, n(i)$. We let

$s_{ij} \equiv$ mean *processing time* for workcenter i required to perform the j th type of operation, $1 \leq i \leq M$, $1 \leq j \leq n(i)$ (by convention, at the L/UL, s_{11} is the *sum* of the load and unload times).

The $\{s_{ij}\}$ must satisfy

$$0 < s_{ij}^- \leq s_{ij} \leq s_{ij}^+, \quad (1)$$

where the limits s_{ij}^\pm are chosen to satisfy the toolmaker specifications to avoid premature tool breakages and to provide acceptable surface quality. Gray et al. (1990) present several references showing how to compute s_{ij}^- and s_{ij}^+ from the geometry of the workpiece and the minimum and maximum tool speeds for a given tool type, workpiece material, and depth-of-cut requirements.

Completed parts exit the FMS at the L/UL station and are immediately replaced by new, fresh parts. This means that the total number of parts (or pallets) in the system is kept at a constant level K , and that a closed queueing network is required to model the facility operations. (K will also be referred to as the work-in-progress (WIP)-level of the system.)

The expected number of times that each part visits workcenter i for the j th type of operation is V_{ij} . The V 's are the so-called *visit ratios*, and $V_{i1} = 1$, since the load/unload operation is performed only once per part. We also use the notation $V_i \equiv \sum_{j=1}^{n(i)} V_{ij}$ to denote the expected total number of times that a fresh part will visit workcenter i .

The mean transport delay (both inbound and outbound) associated with one type j operation at workcenter i is given by D_{ij} . We assume one part-type family and let TH be the desired FMS throughput. Given the total number of pallets K in the system, the minimum $\{s_{ij}^-\}$ and maximum $\{s_{ij}^+\}$ processing times, the transport delays $\{D_{ij}\}$, and visit ratios $\{V_{ij}\}$, it is possible to determine the minimum and maximum throughputs (TH^- and TH^+ , respectively) of which the system is capable. The continuity and decreasing monotonicity of the throughput as a function of the processing times $\{s_{ij}\}$ in approximate mean value analysis (MVA) is given by Schweitzer and Seidmann (1990b). The problem is feasible if and only if the throughput goal TH satisfies

$$TH^- \leq TH \leq TH^+. \quad (2)$$

The last properties are believed true for any exact model as well.

We let $g_{ij}(s_{ij})$ denote the expected cost of performing *one* type- j operation at workcenter i , for $1 \leq i \leq M$, $1 \leq j \leq n(i)$, if the processing time is s_{ij} (where $s_{ij}^- \leq s_{ij} \leq s_{ij}^+$). We assume that the tool-cost functions $g_{ij}(s_{ij})$ are known and satisfy, for $s_{ij}^- \leq s_{ij} \leq s_{ij}^+$,

$$g_{ij}(s_{ij}) > 0, Dg_{ij}(s_{ij}) < 0, D^2g_{ij}(s_{ij}) > 0,$$

where Dg is the first derivative of $g(\cdot)$ and D^2g is the second. Appendix C outlines the derivation of $g_{ij}(s_{ij})$. These common conditions on the first two derivatives of $g_{ij}(s_{ij})$ are based on previous studies (Hax and Candea 1984; Johnson and Montgomery 1974). When Taylor's law relates tool life to tool speed (Drozda and Wick 1983; Olberg et al. 1976; Taylor 1907), Schweitzer and Seidmann (1988) have shown that $g_{ij}(s_{ij})$ has the power form

$$g_{ij}(s_{ij}) = c_{ij}(s_{ij})^{-e_{ij}}, \quad (3)$$

where c_{ij} and $e_{ij} > 0$ (see appendix C). The power function (3) is the one used in this study.

When the s_{ij} values have been determined for a given TH goal, we also get the following output parameters, for $1 \leq i \leq M$ and $1 \leq j \leq n(i)$:

λ_{ij} = throughput at workcenter i , measured in the number of type- j operations per unit of time;

w_{ij} = mean sojourn time of parts at workcenter i , either on queue or in service, for one type- j operation;

N_{ij} = mean number of parts at workcenter i , either in queue or in service, for type- j operation;

$$COSTMIN = \sum_{i=1}^M \sum_{j=1}^{n(i)} \lambda_{ij} g_{ij}(s_{ij})$$

= cost per unit time of operating the system; and

$$\tilde{c} = COSTMIN/TH = \text{cost per part produced.}$$

This model assumes reliable machines, available tools as needed, steady-state operations, no blocking, FCFS scheduling, and negligible setup times or change-over delays. The model allows distinct suboperations at a machine, allows probabilistic rework and inspection tasks, incorporates material-handling delays, and (see section 7) allows parallel machines. Since TH is unchanged, we ignore the impact of changes in the processing times $\{s_{ij}\}$ on the direct labor costs associated with loading and unloading parts and with machine supervision. $COSTMIN$ and \tilde{c} include the purchasing, installation, and replacement costs of the tools involved, which will be proportional to the number of tool replacements.

3. The optimization framework

The optimization problem, *minimizing the tool cost for a given throughput goal*, is formulated as follows:

$$\min \left\{ COSTMIN = \sum_{i=1}^M \sum_{j=1}^{n(i)} \lambda_{ij} g_{ij}(s_{ij}) \right\} \quad (4)$$

subject to

$$s_{ij}^- \leq s_{ij} \leq s_{ij}^+ \quad i = 1, 2, \dots, M; j = 1, 2, \dots, n(i) \quad (5)$$

$$\lambda_{ij} = V_{ij} \lambda_{11} \quad i = 1, 2, \dots, M; j = 1, 2, \dots, n(i) \quad (6)$$

$$\lambda_{11} = K / \sum_{i=1}^M \sum_{j=1}^{n(i)} V_{ij} (w_{ij} + D_{ij}) \quad (7)$$

$$w_{ij} = s_{ij} \quad i \in WC_{AS}; j = 1, 2, \dots, n(i) \quad (8)$$

$$w_{ij} = s_{ij} + \frac{K-1}{K} \sum_{t=1}^{n(i)} N_{it} s_{it} \quad i \in WC_{FCFS}; j = 1, 2, \dots, n(i) \quad (9)$$

$$N_{ij} = \lambda_{ij} w_{ij} \quad i = 1, 2, \dots, M; j = 1, 2, \dots, n(i) \quad (10)$$

$$\lambda_{11} = TH \quad (11)$$

$$\text{all } w_{ij}, \lambda_{ij}, N_{ij} \geq 0. \quad (12)$$

Here WC_{AS} and WC_{FCFS} denote the set of workcenters whose service discipline is AS and FCFS, respectively.

The objective function (4) is equivalent to minimizing the *tool wearout and replacement cost per part* \tilde{c} , since the throughput goal TH is fixed. Note that a product-form queueing solution does not exist here, since the mean processing times for the various tasks at each workcenter are not necessarily identical. Constraint (5) gives the technological limitations on the maximum and minimum processing times for each operation. Equation (6) expresses the number of type- j operations per unit time at workcenter i being equal to the L/UL throughput λ_{11} multiplied by the visit ratio V_{ij} . Constraint (7) expresses the throughput λ_{11} at the L/UL station (via Little's law) as the part population K divided by the mean production flow time.

Equation (8) equates sojourn times with service times for ample servers. Constraint (9) is the usual MVA approximation for FCFS servers (Schweitzer 1979). It expresses the mean sojourn time w_{ij} at an FCFS server as the sum of the mean service time s_{ij} and the mean queueing time. Equation (10) is Little's law, while equation (11) expresses our goal that the FMS throughput at L/UL equals the target value TH . The algorithm for solving the optimization problem (4)–(12) has been developed by Schweitzer and Seidmann (1989).

4. The base case

The set of decision parameters used in managing the FMS performance will be illustrated using the base-case data set defined in this section. This data set represents a metal-cutting FMS with one L/UL station ($i = 1$), six machining centers ($i = 2, \dots, 7$), one contour-measurement machine ($i = 9$), and two part-cleaning and part-washing stations ($i = 8, 10$). Only the processing rates at the six machining centers can be altered. The FMS operates with $27(=K)$ pallets, and the required throughput target is 0.055 parts/minute ($=TH$). Table 1 depicts the visit ratios, the transporter delays, the processing time ranges, and the parameters of the tool-cost function for the base case used throughout this article. The processing rates at centers $i = 1, 8, 9, 10$ are constant, since these centers do not involve machining operations. For this reason, the cost of operating these centers is omitted from $COSTMIN$. Throughout the article, we exhibit only one example per case in order to supply detailed *insights* on the underlying dynamic phenomenon. All examples use this base case or minor variants.

Nominal values of s_{ij} are in practice chosen by handbook, tool-manufacturer recommendation, historical precedent, or by a one-machine economic model (Drozda and Wick 1983) that balances direct machine-operating cost per hour against tool cost (assuming 100% utilization, no starvation, etc.). The nominal values of s_{ij} in table 2 were chosen by using a one-machine minimum-cost cutting-speed model for iron-base alloy with carbide tools, as in Drozda and Wick (1983). Inserting the nominal s_{ij} values into the closed-queueing-network model leads to a nominal throughput of $TH = 0.055$ parts/minute. In the next

Table 1. Process requirements and tool-cost functions for the base case.

Machine i	Service discipline	Operation j	V_{ij}	D_{ij} (min)	s_{ij}^- (min)	s_{ij}^+ (min)	c_{ij}	e_{ij}
1	FCFS	1	1.0	14.3	6.0	6.0	—	—
2	FCFS	1	1.0	7.0	1.5	4.2	791	1.76
3	FCFS	1	0.50	6.0	3.8	6.1	2471	2.739
3	FCFS	2	0.70	6.0	7.2	8.4	1577	2.17
3	FCFS	3	0.40	10.0	1.6	4.0	2018	1.68
4	FCFS	1	0.50	6.0	3.8	6.1	2471	2.739
5	FCFS	1	1.0	9.0	5.5	8.5	3720	2.80
5	FCFS	2	0.50	2.0	7.0	12.9	2431	2.80
6	FCFS	1	0.70	5.0	17.0	29.0	5198	2.02
7	FCFS	1	0.70	5.0	4.5	11.0	5198	2.02
7	FCFS	2	0.30	5.0	13.0	23.1	4741	3.21
8	AS	1	3.0	12.0	5.4	5.4	—	—
9	FCFS	1	0.15	4.9	27.0	27.0	—	—
10	AS	1	1.0	10.0	5.4	5.4	—	—

Notes: V_{ij} = visit ratio; D_{ij} = transporter time (min); $s_{ij}^{+(-)}$ = longest (shortest) feasible processing times (min); c_{ij} and e_{ij} denote the cost proportionality constant and the exponent, respectively. Here $K = 27$, $TH = 0.055$ parts/minute, $TH^- = 0.048$ parts/minute, $TH^+ = 0.080$ parts/minute.

Table 2. Nominal and optimized system performance.

FMS performance									
Machine (i)	Visit (j)	Using nominal processing times				Using optimized processing times			
		s_{ij} (min)	W_{ij} (min)	U_i	$\Sigma_j V_{ij} g_{ij}$	s_{ij} (min)	W_{ij} (min)	U_i	$\Sigma_j V_{ij} g_{ij}$
1	1	6.0	8.83	0.330	0	6.0	8.79	0.330	0
2	1	2.3	2.62	0.127	182.61	4.2	5.4	0.231	63.28
3	1	4.4	9.79	0.473	285.71	6.1	15.01	0.579	98.24
	2	7.9	13.29			8.4	17.31		
	3	2.0	7.39			4.0	12.91		
4	1	4.4	4.99	0.122	21.35	6.1	7.26	0.168	8.73
5	1	6.4	21.25	0.660	22.05	8.5	48.04	0.822	10.24
	2	11.0	25.85			12.9	52.44		
6	1	25.0	379.73	0.970	5.46	24.62	281.88	0.948	5.60
7	1	6.4	24.41	0.581	85.69	11.00	43.95	0.727	28.81
	2	20.0	38.01			18.41	51.37		
8*	1	5.4	5.40	0.898	0	5.4	5.40	0.898	0
9	1	27.0	34.45	0.225	0	27.0	34.45	0.225	0
10*	1	5.4	5.40	0.297	0	5.4	5.40	0.297	0
Cost/part:				\$602.87			\$214.90		

*Indicates the ample server machine (or delay) where "utilization" is interpreted as the expected number of parts in process.

Note: $\Sigma_j V_{ij} g_{ij}$ denotes the cost contribution of machine i to each part, and U_i is the machine utilization. Here $K = 27$ pallets and $TH = 0.055$ parts/minute.

step, the processing times were optimized (in table 2) subject to achieving the same throughput goal $TH = 0.055$ parts/minute.

Considerable cost savings are demonstrated here (a threefold savings in tool cost implies a 5%–10% reduction in the total FMS operating costs), even though the throughput goal $TH = 0.055$ parts/minute is close to the minimum throughput ($TH^- = 0.048$ parts/minute) of the FMS. The key to achieving these cost savings is the fact that the bottleneck workcenter (machine 6) was speeded up while all the other workcenters were slowed down. Note that, after the optimization, s_{71} goes *down* while s_{72} goes *up*. A static one-machine model would not predict this kind of subtle behavior, which depends on the relative slopes of the cost curves $g_{ij}(s_{ij})$.

One-machine models are unsatisfactory because they ignore starvation and throughput goals (so stations cannot be operated at speed with lowest cost per operation), and they lead to unbalanced systems.

Table 2 indicates that it is not optimal to seek either equal utilization or equal sojourn time or equal queue length at every workcenter, or equal cost contribution from every workcenter. It is also not necessarily desirable to reduce the s_{ij} 's times in order to compensate for large transporter delays D_{ij} 's.

5. Parts routing and assignment of tasks to machines

Part routing and the assignment of tasks to machines are problems that have received considerable attention in the literature due to their significant impact on FMS performance (Gray et al. 1990). Throughout these studies, however, processing rates are assumed to be constant, and queueing delays are ignored. Such routing problems become very important in FMSs when the replacement of people by machines eliminates the informal system for monitoring and correcting unbalanced workloads.

Our purpose in this section is to address these lacunae in the management of FMSs and to integrate, within the mean value analysis of closed queueing networks, the basic attributes of part-routing decisions. We introduce part-routing procedures for maximizing the upper bound of the system capacity (TH^+) and for reducing the operating cost ($COSTMIN$), assuming that the values of K and D_{ij} are unchanged.

We start our exposition with a case where a given operation can be split between being performed at two workcenters: as operation *c* on workcenter *a*, with time s_{ac} and cost $g_{ac}(s_{ac})$, or as operation *e* on workcenter *d*, with time s_{de} and cost $g_{de}(s_{de})$. Since the amount of work to be performed per part type is given, we set $V_{ac} + V_{de} = \text{Constant}$, and want to find the optimal value of V_{ac} . Direct analytical evaluation of the optimal V_{ac} value seems intractable. Instead, we find the optimal routing by a univariate search. This search is facilitated by evaluating the following *constrained derivative* (reduced gradient) (Himmelblau 1972), which treats only V_{ac} as an independent decision variable and V_{de} as dependent:

$$\frac{\delta COSTMIN}{\delta V_{ac}} = \frac{\partial COSTMIN}{\partial V_{ac}} - \frac{\partial COSTMIN}{\partial V_{de}}, \quad (13)$$

the symbol δ denoting the constrained derivative. For evaluating the right-hand side of equation (13), we use the following theorem.

Theorem 1. The partial derivative of the optimal cost value ($COSTMIN$) with respect to the part routes V_{ac} is given by

$$\begin{aligned} \frac{\partial COSTMIN}{\partial V_{ac}} &= THg_{ac}(s_{ac}) + b \frac{TH}{K} [s_{ac} + D_{ac}] \\ &+ \frac{b(TH)^2(K-1)}{K^2} \frac{V_a s_{ac}^2 + \sum_{j=1}^{n(a)} V_{aj} s_{aj}^2}{1 - \frac{K-1}{K} TH \sum_{j=1}^{n(a)} V_{aj} s_{aj}} \\ &+ \frac{b(TH)^3(K-1)^2}{K^3} \frac{V_a s_{ac} \sum_{j=1}^{n(a)} V_{aj} s_{aj}^2}{\left[1 - \frac{K-1}{K} TH \sum_{j=1}^{n(a)} V_{aj} s_{aj}\right]^2} \end{aligned} \quad (14)$$

where b is the Lagrangian multiplier as given in appendix A, equation (A.7). (If station a is AS, then the two terms involving V_a are set equal to zero.) A similar result holds for $\partial COSTMIN/\partial V_{de}$.

Proof. See appendix A and Schweitzer and Seidmann (1990b).

Performing a line search for the optimal value of V_{ac} , we increase (decrease) V_{ac} if $\delta COSTMIN/\delta V_{ac}$ is negative (positive). Having this closed expression is particularly useful in the case of a shallow minimum, as illustrated next.

A series of load-shifting calculations were performed to determine the benefits of parts reroutes by moving operations from a heavily utilized workcenter to a more lightly utilized workcenter capable of performing the same operation. The example used here is similar to the base case used in all other sections, except that we assume a throughput goal of $TH = 0.079$ parts per minute. This throughput goal was set higher than the base case in order to illustrate the dramatic effects of parts rerouting decisions in high-stress situations. Recall that the visit ratio to workcenter 6 is $V_{61} = 0.70$. We now study the option of moving some of the operations from the workcenter 6 onto workcenter 4, whereby a (new) second operation is created, with tool-cost function g_{42} in this example taken as identical to the original cost function g_{61} . The optimal amount of off-loading of the operation from workcenter 6 to workcenter 4 is determined by minimizing $COSTMIN$ over the variables V_{61} and V_{42} , subject to $V_{61} + V_{42} = 0.70$ (the original value) and nonnegativity constraints. This is essentially a one-dimensional optimization over V_{61} , using the expression for $\delta COSTMIN/\delta V_{61}$ given by equation (13).

The results of the optimization over V_{61} are shown in table 3 and figures 1 and 2. The benefits of off-loading are large (here we get more than 50% reduction in unit cost) if, as occurs here, the tool being off-loaded was originally running at high speed and had a high sensitivity of cost to speed.

Table 3. The cost per part \tilde{c} (\$) as a function of the visit ratio.

V_{61}	V_{42}	\tilde{c} (\$)	$\frac{\delta COSTMIN}{\delta V_{61}}$
0.7	0	519	82.87
0.65	0.05	257	2.23
0.6	0.1	247	0.45
0.55	0.15	243	0.32
0.5	0.2	242	0.12
0.45	0.25	242	-0.06
0.4	0.3	242	-0.24
0.35	0.35	243	-0.49
0.3	0.4	248	-0.63

Note: Here $V_{61} + V_{42} = 0.7$.

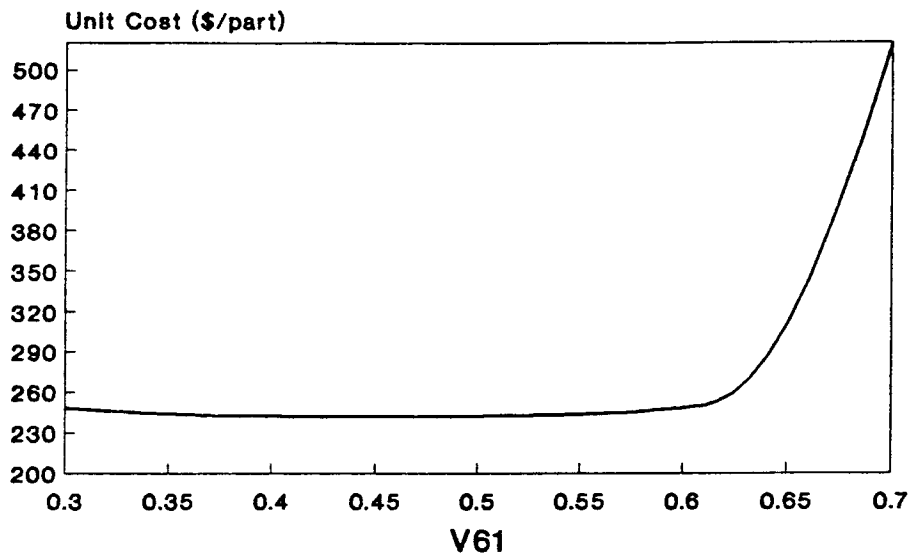


Figure 1. Changes in the cost per part \tilde{c} (\$) as a function of the visit ratio V_{61} . Here $V_{61} + V_{42} = 0.7$; $TH = 0.079$ parts/minute; $K = 27$ pallets.

Note that the optimum is very flat (“shallow bowl” phenomenon), and therefore visit ratios V_{61} anywhere between 0.35 and 0.55 are generally acceptable. Indeed, a search for the “exact” optimum is very difficult without the help of the formula for $\delta COSTMIN/\delta V_{61}$. In this example the minimum cost of \$241.68 is achieved at $V_{61} = 0.468$. The impact of rerouting parts from workcenter 6 (operation 1) to workcenter 4 (operation 2) on the values of the optimal processing times and on the cost contribution of each operation concerning workcenters 4 and 6, as well as on the utilization of these two workcenters, is presented in table 4.

Note from table 4 that the 54% cost savings of $\$519 - \$242 = \$277$ per part do not occur at workcenters 4 and 6 themselves, but as a result of slowing down *other* operations. We also notice from table 4 that s_{41}^* is *slowed down* despite the fact that operations are

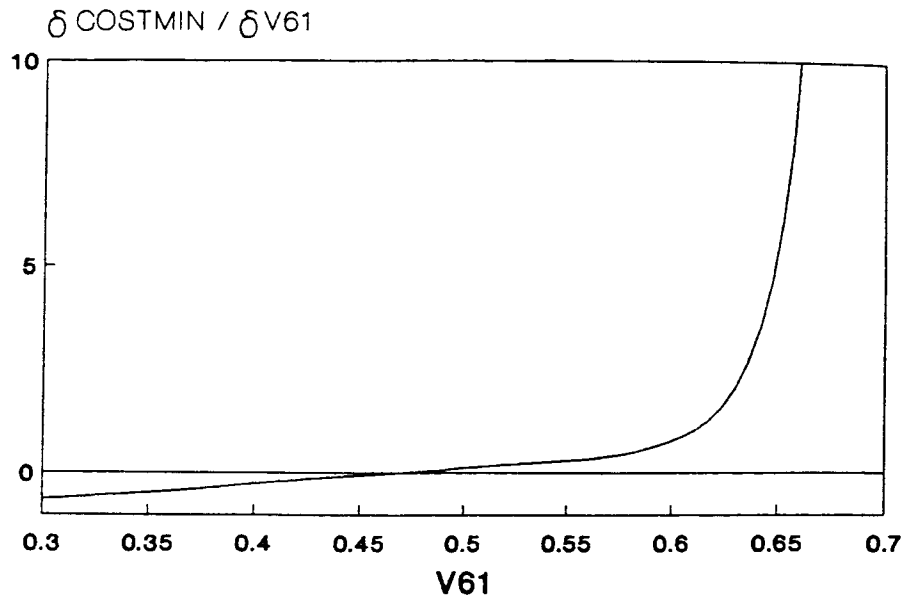


Figure 2. Changes in the derivative $\delta \text{COSTMIN} / \delta V_{61}$ as a function of the visit ratio. Here $V_{61} + V_{42} = 0.7$; $TH = 0.079$ parts/minute; $K = 27$ pallets.

Table 4. The impact of rerouting parts from workcenter 6 (operation 1) to workcenter 4 (operation 2).

V_{61}	V_{42}	s_{61}^* (min)	s_{41}^* (min)	s_{42}^* (min)	U_6	U_4	$V_{61}g_{61}$ \$/part	$V_{41}g_{41}$ \$/part	$V_{42}g_{42}$ \$/part
0.7	0	17.00	3.93	—	0.940	0.155	11.90	29.02	—
0.65	0.05	17.00	6.10	17.00	0.873	0.308	11.04	8.73	0.85
0.6	0.1	17.00	6.10	17.60	0.806	0.380	10.20	8.73	1.58
0.55	0.15	17.00	6.10	18.36	0.739	0.459	9.35	8.73	2.18
0.5	0.2	17.00	6.10	17.46	0.671	0.517	8.50	8.73	3.22
0.45	0.25	17.40	6.10	17.00	0.619	0.577	7.30	8.73	4.25
0.4	0.3	18.40	6.10	17.00	0.521	0.711	4.82	8.73	5.95
0.35	0.35	18.86	6.10	17.00	0.521	0.711	4.82	8.73	5.95
0.3	0.4	19.55	6.10	17.00	0.463	0.778	3.85	8.73	6.80

added to workcenter 4. It is also noteworthy that s_{42}^* is neither convex nor monotone in V_{61} . In general, cost minimization requires *partial off-loading* rather than *total off-loading* of operations: an operation must be split between two or more machines. This will require duplicate tooling and an administrative burden, which may be justified by the potentially large reduction in unit costs.

Rerouting parts among workcenters capable of performing identical tasks also has a direct impact on the maximal system capacity TH^+ . The following theorem helps determine the optimal part routes (optimal assignment of the production load between two workcenters) that maximize the system capacity.

Theorem 2. If a certain production task can be performed on machine **a** as task **b**, or on machine **c** as task **d**, and $V_{ab} + V_{cd} = \text{Constant}$, then the *constrained derivative* of the maximum FMS throughput rate with respect to V_{ab} is given by

$$\frac{\delta TH^+}{\delta V_{ab}} = \frac{\partial TH^+}{\partial V_{ab}} - \frac{\partial TH^+}{\partial V_{cd}} \Big|_{V_{cd} = \text{Constant} - V_{ab}} \quad (15)$$

where

$$\frac{\partial TH^+}{\partial V_{ab}} = - \frac{\partial Q(TH^+, \mathbf{V})}{\partial V_{ab}} \Big/ \frac{\partial Q(TH^+, \mathbf{V})}{\partial TH^+} \quad (16)$$

with

$$\frac{\partial Q(TH^+, \mathbf{V})}{\partial TH^+} = \frac{1}{TH^+} + \frac{TH^+(K-1)}{K^2} \sum_{i \in WC_{FCFS}} V_i \frac{NUM_i}{(DEN_i)^2}, \quad (17)$$

$$\begin{aligned} \frac{\partial Q(TH^+, \mathbf{V})}{\partial V_{ab}} = \frac{TH^+}{K} (s_{ab}^- + D_{ab}) \\ + \frac{(K-1)(TH^+)^2}{K^2 DEN_a} \left\{ NUM_a \left[1 + \frac{K-1}{K} V_a TH^+ s_{ab}^- \right] + V_{ab} (s_{ab}^-)^2 \right\}, \end{aligned} \quad (18)$$

and

$$NUM_i = \sum_{j=1}^{n(i)} V_{ij} (s_{ij}^-)^2, \quad (19)$$

$$DEN_i = 1 - \frac{K-1}{K} TH^+ + \sum_{j=1}^{n(i)} V_{ij} s_{ij}^-. \quad (20)$$

(This formula assumes that machine **a** is FCFS. An expression similar to equation (16) holds for $\partial TH^+ / \partial V_{cd}$.)

Proof. See appendix A.

Using our base case, we investigated the effect of off-loading workcenter 6 (operation 1), which had the highest utilization (see table 2), initially onto a workcenter with a light load (workcenter 4) and later onto a workcenter with a relatively heavy load (workcenter 5). We assume here that operation 1 on machines 4, 5 and 6 is functionally the same. The effects on the capacity of the FMS under these two experiments are depicted in figures 3 and 4, respectively.

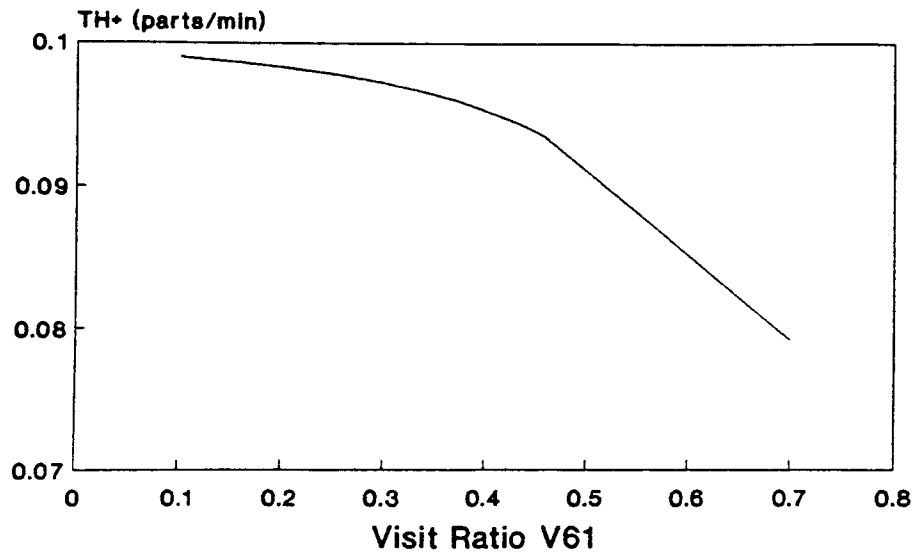


Figure 3. Changes in the maximum capacity TH^+ as a function of the visit ratio. Here $V_{61} + V_{41} = 1.2$; $TH = 0.079$ parts/minute; $K = 27$ pallets.

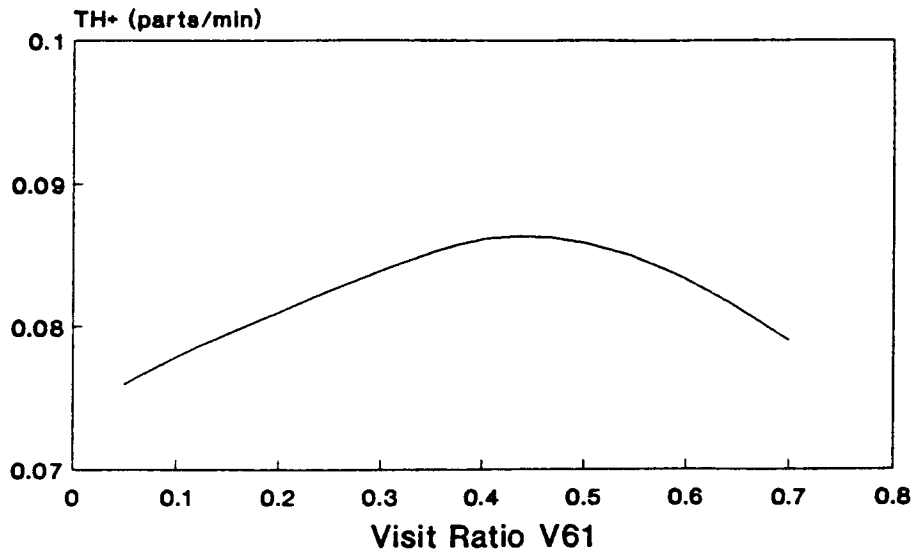


Figure 4. Changes in the maximum capacity TH^+ as a function of the visit ratio. Here $V_{61} + V_{31} = 1.7$; $TH = 0.079$ parts/minute; $K = 27$ pallets.

In figure 3, the maximum capacity is achieved when we move all of the operation from workcenter 6 to workcenter 4. This corner solution results in a capacity increase of 25%. Figure 4 presents the case in which the optimum is an interior solution with a capacity

increase of 14%. In both cases, we again observe the “shallow bowl” phenomenon, which highlights the value of using $\delta TH^+/\delta V_{ab}$ from theorem 2 when searching for the optimal solution.

In general, the case of several simultaneous off-loadings is a multivariate optimization problem. It can be implemented either as a sequence of pairwise off-loadings, as just illustrated, or by a projected gradient search using equations (14) or (16) to compute the components of the gradient.

6. Selection of tool types

Each machining operation can be done with more than a single tool type. For example, steel alloys can be machined by tools made of high speed steel (HSS), tungsten carbide, or aluminum oxide. In practice, tool-type selection for each machining task is done for *one machine at a time* during the process planning phase (Drozda and Wick 1983; Gray et al. 1990; Trucks 1987). This practice ignores the impact of such early tool choices on the overall FMS performance. This factor is important, since selecting a tool type affects the feasible range of processing rates for that particular task ($[s_{ij}^-, s_{ij}^+]$). Consequently, tool-type selection in *one* machine tends to have an immediate effect on the performance of *other* machines.

A lower-cost way of selecting tools is to *simultaneously* consider multiple tooling options for several machines. Enumerating the various tooling combinations and evaluating their minimal cost via the optimization scheme of equations (4)–(12) will result in the overall best set of tools for the prescribed throughput goal TH .

The next example explores the sensitivity of operating costs to tool choice, and examines the interdependence of tool-choice decisions among machines. Two tool options are available at machines 2, 4, and 6. The tool parameters are shown in table 5, with tool I always representing the “base case” (cf. table 1) and tool II denoting another technologically feasible alternative. Tool costs for the same tool vary from one machine to another, since they depend on the nature of the task performed at each machine.

Figures 5a–5c show that neither tool dominates the other over the entire processing time range, i.e., typically one tool is cheaper at higher speeds, while the other one is cheaper at lower speeds. (This situation is typical when one has to select among different tool grades such as HSS, carbide, and zirconium.)

Table 5. Cost coefficients and range of feasible processing times for three pairs of feasible tool types for machines 2, 4, and 6.

Machine (i)	Tool type	Coef. c_{ij}	Exp. e_{ij}	s_{ij}^- (min)	s_{ij}^+ (min)
2	I	791	1.760	1.5	4.2
	II	5450	3.166	2.8	7.1
4	I	2471	2.739	3.8	6.1
	II	5014	3.210	3.5	5.5
6	I	5198	2.020	17.0	29.0
	II	1787	1.670	16.0	26.0

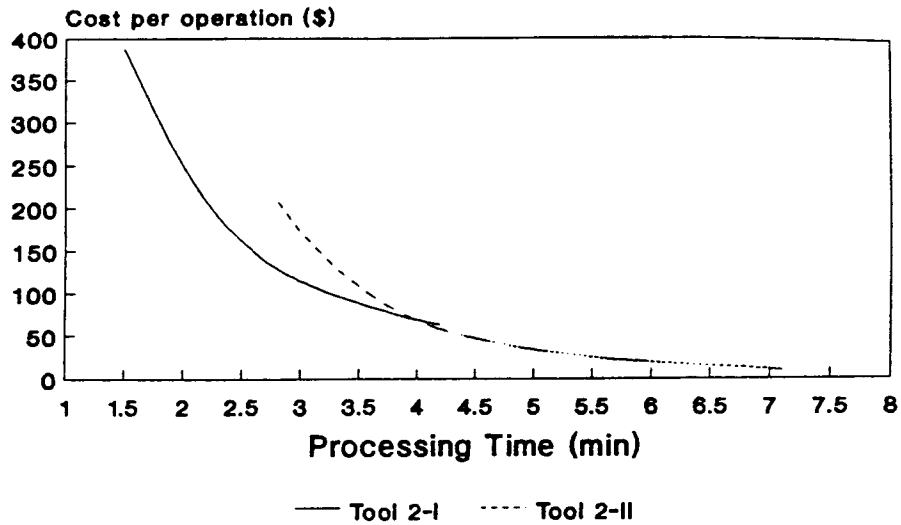


Figure 5a. The two tool costs per operation for machine 2.

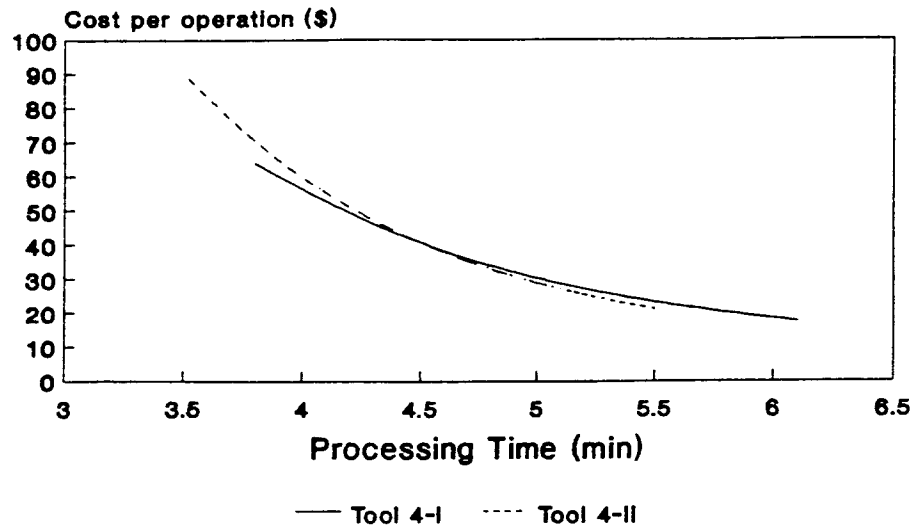


Figure 5b. The two tool costs per operation for machine 4.

The cheaper tool is, of course, preferred when *both* tools are feasible. It is obvious from table 5 data that allowing the option of alternative tools at a given machine may *extend the range of processing times* possible at this machine. At least two possible outcomes must be considered. On the one hand, switching to a higher-speed tool may be preferable, even if it has a higher cost per operation than the base case, if it permits us to reduce tool speeds (and tool costs) at *other* workcenters. On the other hand, switching to a lower-speed tool may be cheaper if it permits operating at a sufficiently lower speed (and cost) than with the original setup. We shall illustrate both possible outcomes next.

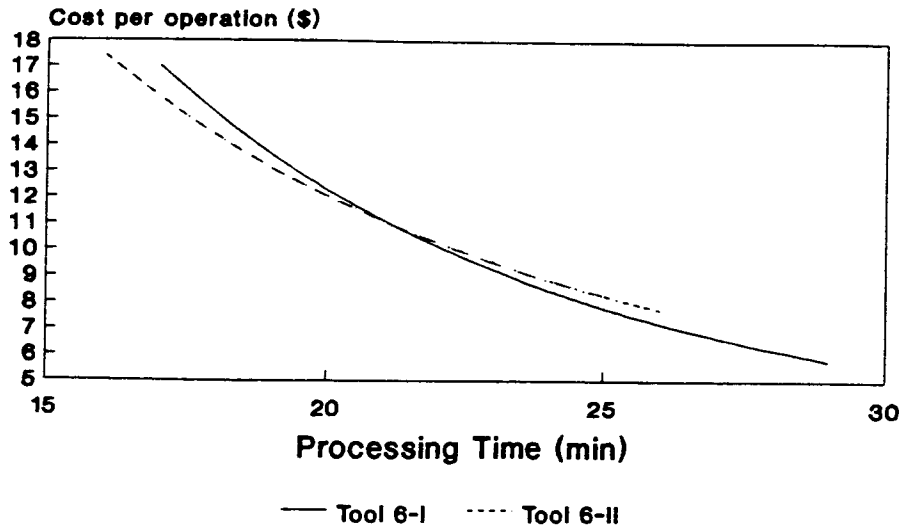


Figure 5c. The two tool costs per operation for machine 6.

Table 6 Complete enumeration of tooling alternatives for $TH = 0.065$ parts/minute and $K = 27$ pallets.

Alternative	Tools used at			s_{21}^* (min)	s_{41}^* (min)	s_{61}^* (min)	\tilde{c} (\$/part)	TH^+ (parts/min.)	TH^- (parts/min.)
	Mach. 2	Mach. 4	Mach. 6						
1	I	I	I	4.2	6.1	19.70	219.21	0.0793	0.0477
2	I	I	II	4.2	6.1	19.57	219.05	0.0831	0.0524
3	I	II	I	4.2	5.5	19.83	220.96	0.0793	0.0477
4	II	I	I	7.1	6.1	19.72	167.05	0.0792	0.0477
5	I	II	II	4.2	5.5	19.58	220.86	0.0831	0.0524
6	II	I	II	7.1	6.1	19.58	166.85	0.0830	0.0523
7	II	II	I	7.1	5.5	19.70	168.87	0.0792	0.0477
8	II	II	II	7.1	5.5	19.43	168.71	0.0830	0.0523

s_{ij}^* denotes the optimal processing time (min.) for task ij . \tilde{c} gives the optimal value of the cost per part for each one of the eight tooling alternatives.

Table 6 displays the results from complete enumeration of all eight tooling alternatives. Each alternative uses the system model (4)–(12) for deriving optimized processing times s_{ij}^* s. It shows that unit costs may be reduced by 25% by appropriate tool choice (the base case in alternative 1 has unit cost $\tilde{c} = 219.21$ \$/part; the best case in alternative 6 has $\tilde{c} = 166.85$ \$/part). Table 6 also shows that *maximal savings require the change of at least two tool types away from the base case*. The least-cost tooling option at alternative 6 is also associated with a 5% increase in the maximal feasible throughput: TH^+ goes up from 0.0793 to 0.0830 parts/minute.

We found that it is very hard to predict beforehand the nature of the tool replacements that should take place. Both tool-replacement possibilities mentioned above are illustrated in this example. The overall change from configuration 1 to configuration 6 is composed of the change of tool at machine 2 from I to II, followed by the change of tool at machine

6 from I to II. The first change, from configuration 1 to configuration 4, reduces costs by *replacing a tool by a slower tool*. (This in fact is responsible for most of the overall benefits.) The second change, from configuration 4 to configuration 6, slightly reduces costs further by *replacing a given tool (I) by a faster one (II)*.

Changes in tool-type choices have extended the range of feasible processing times per operation [s_{ij}^+ to s_{ij}^-]. Consequently, there are cases where switching to a more expensive tool becomes beneficial, since doing so allows us to machine faster on one machine, and thereby to slow down on another, while generating a net saving in the total FMS tool costs. This was noteworthy when contrasted with the conventional one-machine models, which led to inferior tool choices at most of the nonbottleneck machines.

7. Capacity expansion decisions

Earlier studies dealing with capacity increments and machine allocation problems in FMSs assume *constant* processing rates (e.g., Bitran and Tirupati 1988; Boxma et al. 1990; Dallery and Frein 1986; Shanthikumar and Yao 1987, 1988; Vinod and Solberg 1985; Vliet and Rinnooy Kan 1990). This approach ignores the potential for additional savings in tool costs by *reoptimizing* the processing rates as well as *reevaluating* the desired number of pallets K and the tool choices when capacity is added. The major decision problem is how to pick the machine to be supplemented in order to achieve the maximal reduction in tool costs.

We present and evaluate here two heuristic policies that involve a sequence of configuration changes, each time duplicating either the *bottleneck* machine (the machine with highest utilization) or else the machine with the highest contribution to the total cost per part.

Incrementing the number of copies of machine i is approximately modeled by increasing M by 1 and dividing the initial visit ratios V_{ij} equally among the identical copies of the machine. We expect these results to provide an *upper bound* on the value of $COSTMIN$, since server pooling is ignored. Although our model considers only operating costs, it can be easily supplemented by separate estimates of the one-time procurement costs for the machine and associated tooling.

Illustrating these two heuristic policies, we start with the base-case configuration (cf. table 1) with a slightly higher value of TH . Note that machines 1, 8, 9, and 10 did not involve metal-cutting operations and therefore were excluded from consideration. All configurations studied in this section have a higher throughput target $TH = 0.079$ parts per minute to provide a high-stress situation. Since $K = 27$ and TH are unchanged, the part flow times ($F = K/TH$) stay the *same* throughout the capacity expansion process.

The base case is shown as configuration I in table 7. Workcenter 6 is the primary bottleneck (has the highest utilization), and machine 5 is the secondary bottleneck.

Duplication of machine 6 leads to configuration II in table 7. Note that the cost per part has dropped in *half*, while the cost savings at the duplicated machine itself (machine 6) are minimal, going down from \$11.90 to \$8.72 ($= 2 * 4.36$) per part. This phenomenon is due to the fact that reduced congestion at machine 6 allows the other machines to slow down significantly, while still achieving the same throughput goal. (A similar behavior was noted in table 4.) Note also that the new primary bottleneck is machine 7, rather than the secondary bottleneck (machine 5) from configuration I.

Table 7. Capacity expansion as the machines with highest utilization are incremented.

Machine	Configuration					
	I		II		III	
	Util.	Unit cost (\$/part)	Util.	Unit cost (\$/part)	Util.	Unit cost (\$/part)
2	0.261	96.71	0.332	63.28	0.332	63.28
3	0.645	171.37	0.789	100.76*	0.832	98.24*
4	0.155	29.26	0.241	8.73	0.241	8.73
5	0.711	36.67	0.825	21.59	0.876*	17.98
6	0.940*	11.90	0.548	4.36	0.640	3.19
6a	—	—	0.548	4.36	0.640	3.19
7	0.557	174.74*	0.866*	34.48	0.458	14.52
7a	—	—	—	—	0.458	14.52
Total		\$520.65		\$237.55		\$223.64

*Indicates the largest value in each column.

Duplication of bottleneck machine 7 in configuration II now leads to configuration III in table 7. The additional 6% reduction in tool cost per part (i.e., \$14 per part) comes mainly from two sources: slower tool speeds at machine 7 (\$5 per part) and slower tool speeds at machine 5 (\$4 per part). The primary bottleneck in configuration III is machine 5, and coincides with the secondary bottleneck in the previous configuration; this is the usual outcome.

The diminishing marginal benefits in going from configurations I to II and III is evident. This arises because configuration II has two nearly-ties bottlenecks, machines 5 and 7, and upgrading only one of them still leaves the other one active.

A parallel set of configuration changes was performed in which the machine with *highest contribution to the total cost per part* was duplicated. The results of these changes are reported in table 8. By duplicating machine 7 in Configuration I (base case configuration),

Table 8. Capacity expansion as the machines with highest contribution to total cost per part are incremented.

Machine	Configuration					
	I		II'		III'	
	Util.	Unit cost (\$/part)	Util.	Unit cost (\$/part)	Util.	Unit cost (\$/part)
2	0.261	96.71	0.322	66.82	0.332	63.28*
3	0.645	171.37	0.667	134.18*	0.389	50.80*
3a	—	—	—	—	0.389	50.80
4	0.155	29.26	0.185	17.93	0.241	8.73
5	0.711	36.67	0.711	36.67	0.711	36.67
6	0.940*	11.90	0.940*	11.90	0.940*	11.90
7	0.557	174.74*	0.348	35.60	0.458	14.52
7a	—	—	0.348	35.60	0.458	14.52
Total		\$520.65		\$338.71		\$251.21

*Indicates the largest value in each column.

configuration II' is obtained. Configuration III' is the result of duplicating machine 3 in configuration II'. It is apparent from tables 7 and 8 that configuration II' has significantly higher operating cost than configuration II. In other words, choosing which machine to replace via its cost contribution rather than marginal FMS cost contribution can be suboptimal.

Experimenting with several other cases as well, we found that the primary bottleneck keeps shifting as capacity is added, and is usually, but not always, the previous secondary bottleneck. This observation is useful when one wants to consider a migration path for the FMS capacity configuration.

Table 7 also shows that incrementing the number of machines results in *increased* utilization of all machines (except for the one just augmented). The utilization of the machine just augmented decreases (but by less than what would occur if one merely divided the original utilization in half). At first it may appear counterintuitive that "increasing system capacity increases machine utilization." This happens because higher capacity at the bottleneck machine reduces the longer residence times at the bottleneck machines, and therefore permits us to lower tool speeds at other nonbottleneck machines.

Incrementing the original number of machines leads to significant (but diminishing) operational savings. The benefits of machine duplication consist of lower unit costs and higher capacity. These benefits may be greater when incrementing the machine with the *highest utilization* rather than the one with the highest contribution to the total cost. This result may also appear counterintuitive and cannot always be correct, since it ignores both operating costs and procurement costs. It was observed here because the cost contribution per machine ($\sum_{j=1}^{n(i)} V_{ij}g_{ij}(s_{ij})$) was not a good surrogate for ranking marginal costs and/or marginal benefits. The sequence of machine duplications must consider the complex interplay between transporter delays, routing, processing speed, and tool-type selection. It requires a multistep search process and is not obvious.

8. Adjusting the number of pallets (or the WIP level)

In this section, we analyze the impact of changes in the WIP level, or the total number of pallets, on the optimized system performance. Pallets and fixtures are often expensive and may cost thousands of dollars apiece (Solot and Bastos 1988; Stecke 1989; Solot 1990). The conventional wisdom is to minimize the WIP (number of parts or pallets) in the system (Buzacott and Yao 1986). WIP leads to higher inventory carrying costs and greater system lead times. The latter both reduces flexibility to respond to schedule changes and increases the time to detect degradation in quality. These adverse consequences of WIP, while present, *are not the whole story*. We show that *increasing* WIP has the advantages of *increasing* system capacity TH^+ and of *reducing* the unit operating cost \tilde{c} . Both of these beneficial consequences are due, we believe, to the *reduction in machine starvation* when pallets are added. This alleviates the burst behavior in which a machine is alternately starved and working *very fast* on a part (in order to reduce sojourn time and thus meet the system throughput target); less starvation permits us to slow down most machines and thus increase the useful tool life.

These beneficial consequences are apparent from the curves in figures 6, 7, and 8. If the throughput is fixed, say at $TH = 0.065$ parts/minute, then increasing K has a beneficial

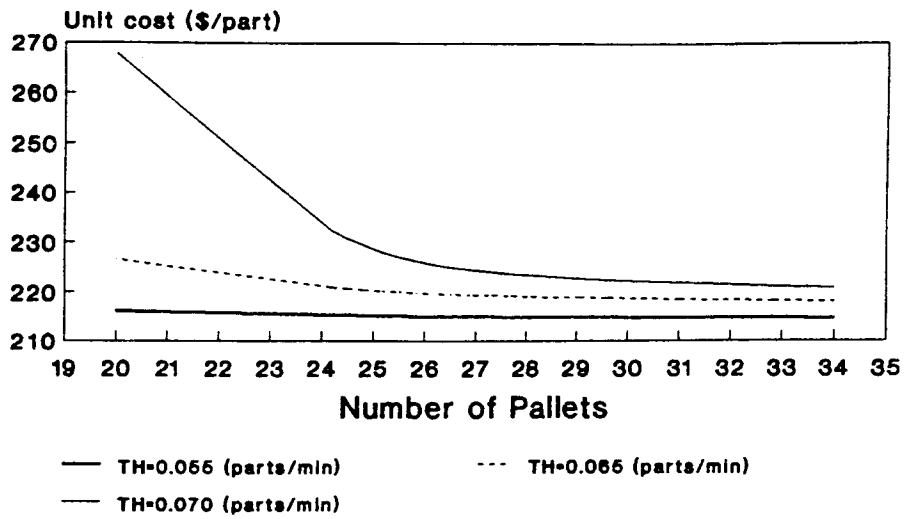


Figure 6. The cost per part \tilde{c} (\$) as a function of the number of pallets.

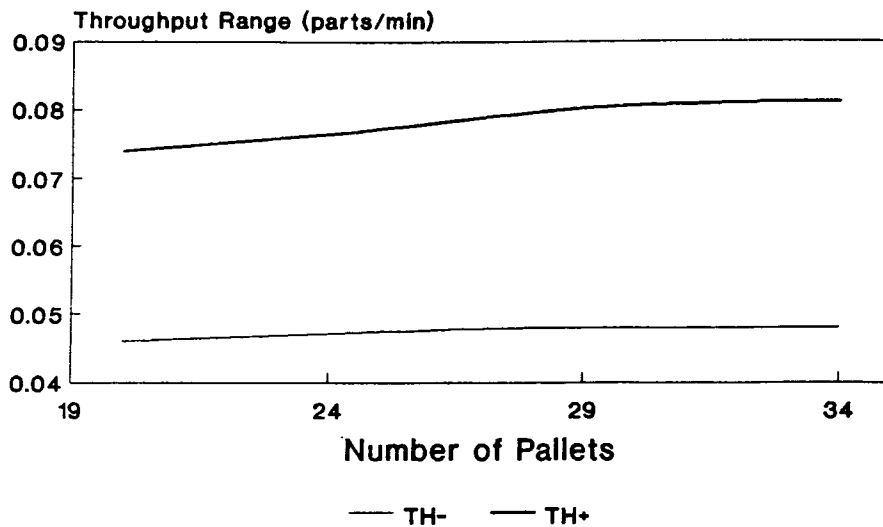


Figure 7. The throughput range $[TH^-, TH^+]$ (parts/minute) as a function of the number of pallets.

effect on the unit cost, as shown in figure 6. The beneficial effect of additional pallets becomes more pronounced at higher throughput levels: increasing the number of pallets from 20 to 34 reduces unit cost by 5% when $TH = 0.065$ parts/minute, and by 30% when $TH = 0.070$ parts/minute. Figure 6 also shows the diminishing marginal benefits of additional pallets.

The maximal system capacity increases from $TH^+ = 0.0736$ to 0.0815 parts/minute when the number of pallets increases from 20 to 34 (figure 7), i.e., one can obtain a 7% increase in the system capacity TH^+ by adding pallets. Diminishing marginal benefits from the increase in K are established here as well. Stated differently, the throughput TH is concave

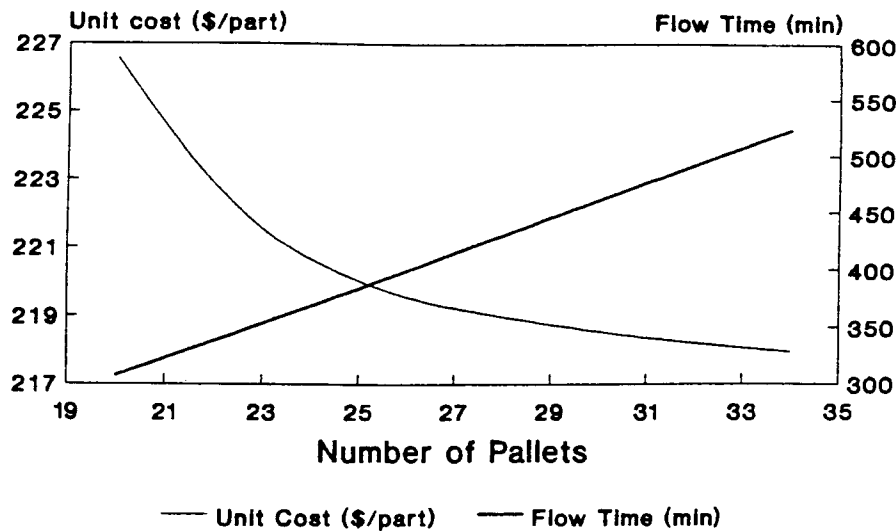


Figure 8. The cost per part \tilde{c} (\$) and the flow time (min) as a function of the number of pallets.

with respect to the number of pallets K . An analytic proof of this property (for the Jackson closed queueing network with *constant* mean processing times) is presented by Shanthikumar and Yao (1988b). It was later applied by Shanthikumar and Yao (1989) for the optimal allocation of the number of pallets (or buffer spaces) in an FMS cell.

Increasing K also has a well-known adverse effect on the FMS flow time $F(= K/TH)$. The effects of increasing the number of pallets on both the cost per part and the flow time are displayed in figure 8. One should also note from this figure the diminishing marginal benefits of adding more and more pallets, and compare these benefits with the constant rate of increase in the flow times.

It should be noted that *machine utilization goes up as K increases*, because we are able to slow down the machines and yet meet the throughput goal TH with less hindrance from starvation (see machines 5, 6, and 7 in table 9).

The above discussion shows that increasing K has *both adverse* (lead time or flow time) and *beneficial* (capacity and unit cost) *consequences*. These beneficial effects increase with the throughput goal for the FMS. A key managerial decision is therefore how to assess the tradeoff between the two effects, while taking into account a one-time pallet procurement cost versus a long-term set of performance implications. This assessment must be sensitive to the length of the planning horizon, and, in the case of small runs, the ability to salvage pallets, fixtures and toolholders for other part types.

9. Sensitivity to throughput goals

Changes in the demand pattern lead to changes in the master production schedule and in the FMS throughput goal TH . We can see (figure 9) that *the tool cost per part \tilde{c} is not constant*. It rises sharply with TH due to the accelerated tool wear. The rate of increase

Table 9. The impact on workcenter utilization of changing the number of pallets (K).

Machine	Machine utilization							
	Number of pallets K							
	20	22	24	26	28	30	32	34
i								
1	0.390	0.390	0.390	0.390	0.390	0.390	0.390	0.390
2	0.273	0.273	0.273	0.273	0.273	0.273	0.273	0.273
3	0.684	0.684	0.684	0.684	0.684	0.684	0.684	0.684
4	0.198	0.198	0.198	0.198	0.198	0.198	0.198	0.198
5	0.761	0.826	0.853	0.879	0.880	0.895	0.905	0.914
6	0.773	0.816	0.864	0.888	0.912	0.924	0.932	0.940
7	0.754	0.754	0.754	0.754	0.754	0.762	0.779	0.790
8*	1.053	1.053	1.053	1.053	1.053	1.053	1.053	1.053
9	0.263	0.263	0.263	0.263	0.263	0.263	0.263	0.263
10*	0.351	0.351	0.351	0.351	0.351	0.351	0.351	0.351

*Indicates the ample server machine (or delay) where "utilization" is to be interpreted as the expected number of parts in process.

Note: Here $TH = 0.065$ parts/minute.

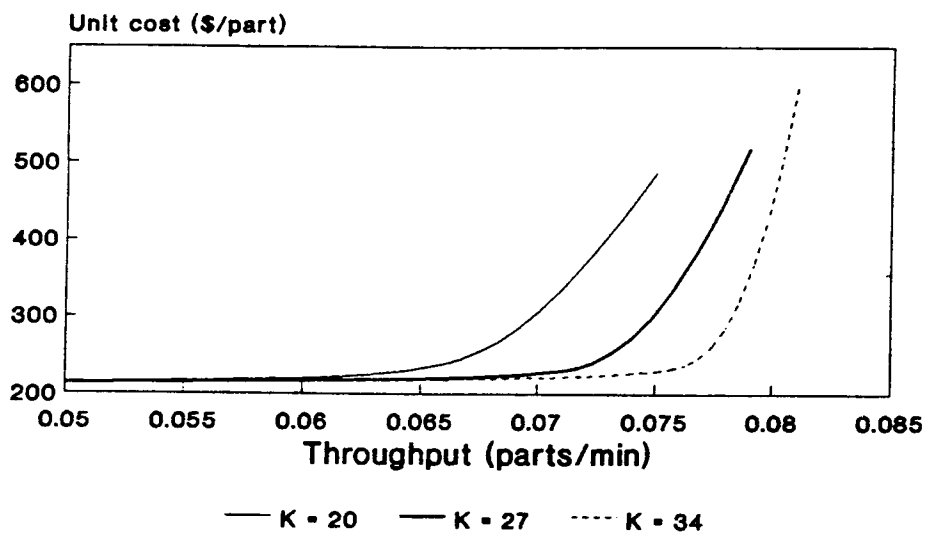


Figure 9. The cost per part \tilde{c} (\$) as a function of the throughput goal (TH), and the number of pallets (K).

gets higher as TH approaches the value of TH^+ . Figure 9 also indicates that increasing the number of pallets results in extending the "flat" portion of the cost function. Furthermore, the structure of these cost functions implies that *management should pay special attention to changes in production goals, even in FMSs*. These observations motivated us to study the multiperiod production planning problem, which is addressed next.

10. Multiperiod production planning

In this section, we are concerned with planning production over a future interval of time, called the *planning horizon*. During this planning horizon, the rate of demand for the products can vary. The purpose of the multiperiod production planning effort discussed here is to map out a program of outputs to meet future customer demands under given constraints, thus utilizing the available manufacturing resources so as to attain high levels of performance criteria. An examination of the production management literature suggests that numerous objectives are often considered by managers and analysts (Eilon 1971; Hax and Candea 1984; McLeavey and Narashiman 1985; Nahmias 1989; Peterson 1971; Silver 1967). Some of these criteria include minimizing costs, maximizing profits, maximizing the net present value of the firm, minimizing lead times, meeting due dates, checking productive feasibility, or maximizing regularity.

Of these various objectives, cost minimization is probably the most prevalent. First, this criterion embraces more specific objectives, such as minimization of inventory levels and variable operating costs. Second, it pertains directly to decision in the production environment (whereas profit is greatly affected by pricing, marketing, and taxation considerations). The last two criteria in the preceding set are also important in practice. The feasibility criterion does not involve optimization. It is a satisficing one, which only requires that an acceptable schedule is found. The regularity criterion presented above reflects the priority given by production managers to minimizing disturbances to the smooth flow of parts.

Several policies for multiperiod production planning in FMSs are presented and examined here. In developing these policies, we divide the planning horizon into T identical weekly time intervals. The demand rate in week t , $1 \leq t \leq T$, is denoted by D_t and may be known or unknown in advance.

The FMS studied here machines a given mix of parts to be used by an assembly operation later on. The output mix of the machined subassemblies must stay the same, but the aggregate overall volume D_t varies with time. In such a case, the part-mix requirements are stationary. The policies studied here simultaneously modify the FMS production goal per unit time (TH_t) for each period t and the processing rates on the individual machines (s_{ij}) for each period. The modifications are chosen to minimize the operating costs while meeting both the varying external customer demand and the technological constraints on the feasible cutting speeds for each task.

The first policy examined here follows a "chase strategy": each week's production equals the week's demand ($TH_t = D_t$). This approach is most flexible, and is used whenever one cannot forecast future demands. It becomes infeasible if any week's demand exceeds the FMS production capacity.

The second policy aims at reducing production irregularities. It assumes that we can forecast demand only *one* period into the future. We then produce at the average demand level for two weeks, after first attempting to meet demand from inventory. The throughput target for week t is given by $TH_t = (D_t + D_{t+1} - I_{t-1})/2$ parts per week, assuming that this is feasible. Here I_{t-1} denotes the inventory at the end of week $t - 1$.

The third policy uses *level production* at rate $TH_t = \bar{D}$, where

$$\bar{D} = \frac{1}{T} \left[\sum_{t=1}^T D_t - I_0 \right]. \quad (21)$$

It assumes that we can forecast the mean net weekly demand rate \bar{D} for the next T weeks. It also assumes that level production is feasible, i.e., that the following conditions hold:

$$I_0 + j\bar{D} \geq \sum_{t=1}^j D_t, \quad j = 1, \dots, T; \quad (22)$$

$$TH^- \leq \bar{D} \leq TH^+. \quad (23)$$

Condition (22) states that the cumulative production capacity for the first j weeks should meet the cumulative demand during these j weeks. The level production policy is supported by the following theorem:

Theorem 3. If the inventory holding costs are negligible and the tool cost function is the same for all periods (i.e., no discounting, and all periods have the same length), then a level production policy at a constant rate of

$$TH_t = \bar{D}, \quad t = 1, 2, \dots, T, \quad (24)$$

parts per period is optimal, provided feasibility conditions (22) and (23) hold.

Proof. See appendix B.

The explanation for this surprising result is that, due to accelerated tool wearout at higher speeds, any deviation from level production will cost more in the *higher-speed* periods than it may save in the *lower-speed* periods.

The major advantage of this policy is its simplicity, as well as its achievement of two major objectives: minimizing production costs along with minimizing the production disturbances.

Finally, we address a fourth policy based on the *exact optimization* of each week's production. This policy assumes that one can forecast all T weeks' demands perfectly. The optimization problem formulation, including both tool costs and inventory holding costs, is

$$\min \left[\sum_{t=1}^T \text{COSTMIN}_t(TH_t) + \sum_{t=1}^T h_t I_t \right] \quad (25)$$

subject to

$$I_t \geq 0, \quad 1 \leq t \leq T \quad (\text{no backlogging}) \quad (26)$$

$$TH^- \leq TH_t \leq TH^+, \quad 1 \leq t \leq T \quad (\text{FMS capacity constraint}) \quad (27)$$

$$I_t = I_{t-1} + TH_t - D_t, \quad 1 \leq t \leq T, \quad (\text{inventory balance}) \quad (28)$$

where

T = planning horizon (number of weeks)

h_i = inventory holding cost (\$/part/week) for week i

$COSTMIN_t(TH_t)$ = minimum tool cost for TH_t parts/week in week t

TH_t = production rate per week (for week t)

I_t = inventory level at the end of week t ($I_0 \geq 0$ is given)

Problem (25)–(28) is minimizing a convex, separable objective function subject to linear constraints. The derivatives of the objective function $COSTMIN_t(TH_t)$ are known from Schweitzer and Seidmann (1989). The functions $COSTMIN_t(TH_t)$ increase monotonically with TH_t . These properties can be used in simplifying the solution of problem (25)–(28). For example, one can fit a polynomial to the values of $COSTMIN_t$ as a function of TH_t and then use this approximation function explicitly in expression (25). A standard software code can then be used to perform the nonlinear optimization. Once the desired values of TH_t are given, it is possible to get the optimal processing costs using equations (4)–(12). Alternatively, a piecewise linear approximation to $COSTMIN_t(TH_t)$ is also feasible.

Note: The constraint $TH_t \geq TH^-$ may not be binding, if one can *shut down* the FMS cell for part of the week, and run at a feasible speed the rest of the week. So for any $TH_t \leq TH^-$, the least expensive policy is to run at (minimal) speed TH^- for a fraction TH_t/TH^- of the week and remain idle the rest, with a tool replacement cost of

$$COSTMIN_t(TH_t) = \frac{TH_t}{TH^-} COSTMIN_t(TH^-), \quad 0 \leq TH_t \leq TH^-. \quad (29)$$

Comparative evaluation of the first three policies outlined above is discussed next. The empirical study consisted of running the FMS base case from section 4 for $T =$ six consecutive weeks in order to satisfy the weekly demands outlined in the second column of table 10.

The average demand is $\bar{D} = 700$ parts per week. With an operation running 24 hours per day, 7 days per week, this corresponds to an average production rate of 0.069 parts/minute, which is within the FMS capacity range of [$TH^+ = 0.080$, $TH^- = 0.048$] parts/minute. The coefficient of variation C_v of the weekly demand is 11.7%. Initial inventory level is zero parts ($I_0 = 0$). The number of pallets assumed here is $K = 28$. The rest of the data items are given in tables 1 and 2 above.

Using the first three policies discussed above, we evaluated three different production schedules to meet the six-weeks market demand. We focus on the tool replacement costs by setting the inventory holding costs to be $h_t = 0$, $t = 1, \dots, 6$.

Table 10. The impact of three alternate multiperiod production targeting schemes on the finished goods inventory and on the total tool costs.

Week	Demand	Plan 1		Plan 2		Plan 3	
		Prod. level	End inv.	Prod. level	End inv.	Prod. level	End inv.
1	600	600	0	650	50	700	100
2	700	700	0	725	75	700	100
3	800	800	0	725	0	700	0
4	600	600	0	650	50	700	100
5	700	700	0	725	75	700	100
6	800	800	0	725	0	700	0
C_v	0.117	0.117		0.05		0	
Avg. monthly inventory		0		41.66		66.66	
Total tool costs		\$1,402,710		\$938,665		\$933,282	

Notes: Plan 1 follows the period-by-period demand; Plan 2 looks ahead for one-period moving average; Plan 3 levels the production.

The results presented in table 10 show that the first policy has the highest cost (\$1,402,710). This implies that *production volume flexibility comes at a high price*. The second policy reduced the costs by 40%, since the coefficient of variation in the weekly production volume dropped from 11.7% to 5%. The third policy presents the optimal plan (cf. theorem 3). This approach is most desirable when one can forecast the future demand profile and when level production is feasible. In our particular example, the optimal total cost is only modestly better than under the second policy, which requires forecasting only one period into the future. This example shows that *considerable cost savings are possible by production smoothing*. It also shows that *even a modest amount of myopic demand smoothing can have major benefits*.

11. Conclusions

This article has dealt with several decision problems associated with performance management in FMSs. Two salient features of the problems discussed are that the processing rates of the machining centers can be revised and that the system has to meet a given throughput goal per unit time. The interaction effects between a large set of managerial decision options and the system performance are studied using both analytical and numerical methods. The decision options studied here include part routing, the WIP levels, capacity expansions, tool type selections, throughput goals, and multiperiod production planning.

These decision options interact with each other in a complicated way. Good performance management requires a *systems* model; decisions cannot be made at a component (i.e., tool or machine) level. Systems models must integrate features of queueing, optimization,

tool technology, and manufacturing costs. Ultimately we would anticipate these models running in real time whenever there is a change in the system setup or in the master production schedule.

We have shown that the ability to adjust system behavior (tool speeds, WIP level, parts routes, etc.) due to changing operating conditions can lead to some *unexpected* responses: 1) increasing WIP *saves* money; 2) increasing capacity *increases* machine utilizations; 3) using more expensive tools can be better; 4) having high variability in machine utilization may be beneficial; and 5) even small amounts of production irregularity can have a large impact on the operating costs. Other unexpected (and perhaps counterintuitive) system responses include changes of two processing times on the same machine in *opposite* directions, and the fact that adding operations to a machine may lead to an *increase* in the processing times of tasks previously assigned to that machine. *This demonstrates the need for a complete reevaluation of performance management when performance tuning is allowed, including a critical reassessment of many conventional conclusions.*

For the case studies reported here, we found that the major cost savings, in order of decreasing importance, are

1. Tool-speed selection
2. Part rerouting and assignment of tasks to machines
3. Tool-type selection
4. Setting multiperiod throughput goals
5. Adjusting the number of pallets

All the above decision options are relatively easy to implement and should lead to significant improvements in many facilities. The nature of these performance tuning actions is such that *some decisions must be made at the run time* and cannot be made as a part of the facility design or the process plan. The reason is that we found very close coupling among most of the system parameters, e.g., increasing the number of machines in one workcenter affects the mean utilization at all other workcenters, or changing the WIP level affects tool costs and the range of feasible throughput goals. *Other decisions studied here have long-range strategic implications*, e.g., setting nonlinear marginal transfer prices and setting multiperiod goals for the master production scheduling.

The investment needed for improved performance management and tuning in FMSs is relatively small in comparison with the high capital investment costs needed for their installation and maintenance (Suri 1988). Our results provide a strong incentive for such an effort, which should be both practical and profitable.

Acknowledgment

Partial support for this study has been provided by the Center for Manufacturing and Operations Management (CMOM) at the William E. Simon Graduate School of Business Administration, University of Rochester. This article benefited from useful comments by the referees, an anonymous associate editor, and Professor K.E. Stecke of the University of Michigan.

Appendix A. Proofs of theorems 1 and 2

Proof of theorem 1

Schweitzer and Seidmann (1989) present the following problem reformulation:

$$COSTMIN = \min_{s_{ij}} \left\{ TH \sum_{i=1}^M \sum_{j=1}^{n(i)} V_{ij} g_{ij}(s_{ij}) \right\} \quad (A.1)$$

subject to

$$s_{ij}^- \leq s_{ij} \leq s_{ij}^+; 1 \leq i \leq M; 1 \leq j \leq n(i) \quad (A.2)$$

$$\frac{TH}{K} \left\{ \sum_{i=1}^M \sum_{j=1}^{n(i)} V_{ij}(s_{ij} + D_{ij}) + \sum_{i \in WC_{FCFS}} V_i \frac{\frac{K-1}{K} TH \sum_{j=1}^{n(i)} V_{ij}(s_{ij})^2}{1 - \frac{K-1}{K} TH \sum_{t=1}^{n(i)} V_{it} s_{it}} \right\} - 1 = 0 \quad (A.3)$$

$$\frac{K-1}{K} TH \left[\max_{i \in WC_{FCFS}} \sum_{j=1}^{n(i)} V_{ij} s_{ij} \right] < 1. \quad (A.4)$$

We rewrite this reformulation as

$$COSTMIN = \min_{\mathbf{S}} \{A(\mathbf{S}, \mathbf{V})\} \quad (A.5)$$

subject to

$$B(\mathbf{S}, \mathbf{V}) = \mathbf{0} \quad (A.6)$$

and the inequality constraints (A.2) and (A.4), where $\mathbf{S} = [s_{ij}]$ and $\mathbf{V} = [V_{ij}]$.

Let b denote the Lagrangian multiplier associated with constraint (A.6). The Lagrangian function is given by

$$L(\mathbf{S}, \mathbf{V}, b) = A(\mathbf{S}, \mathbf{V}) - bB(\mathbf{S}, \mathbf{V}). \quad (A.7)$$

For any fixed value of $\mathbf{V} = [V_{ij}]$, let $\mathbf{S}^*(\mathbf{V})$ and $b^*(\mathbf{V})$ denote the associated optimum, assumed unique and differentiable. (These assumptions will hold except for certain break-points where only directional derivatives exist.)

Then the desired derivatives are given by (see below)

$$\frac{\partial COSTMIN}{\partial V_{ac}} = \frac{\partial L(\mathbf{S}, \mathbf{V}, b)}{\partial V_{ac}} \Big|_{\mathbf{S}=\mathbf{S}^*(\mathbf{V}); b=b^*(\mathbf{V})} \quad (\text{A.8})$$

However, when a given operation can be split between operation c on machine a or operation e on machine d , we have

$$V_{ac} + V_{de} = \text{Constant} \quad (\text{A.9})$$

To find the best split between a and d , we need to compute the *constrained derivative*

$$\frac{\partial COSTMIN(\mathbf{V})}{\partial V_{ac}} = \frac{\partial COSTMIN(\mathbf{V})}{\partial V_{ac}} - \frac{\partial COSTMIN(\mathbf{V})}{\partial V_{de}}. \quad (\text{A.10})$$

The right-hand side of equation (A.10) is evaluated from equation (A.8).

The derivation of equation (A.8) is based upon

$$COSTMIN(\mathbf{V}) = L(\mathbf{S}, \mathbf{V}, b) \Big|_{\mathbf{S}=\mathbf{S}^*(\mathbf{V}); b=b^*(\mathbf{V})}, \quad (\text{A.11})$$

which immediately follows from the optimality conditions:

$$COSTMIN(\mathbf{V}) = A(\mathbf{S}^*(\mathbf{V}), \mathbf{V}) \quad (\text{A.12})$$

and

$$B(\mathbf{S}^*(\mathbf{V}), \mathbf{V}) = \mathbf{0}. \quad (\text{A.13})$$

Differentiating equation (A.11) with respect to V_{ac} gives

$$\begin{aligned} \frac{\partial COSTMIN(\mathbf{V})}{\partial V_{ac}} &= \sum_{i=1}^M \sum_{j=1}^{n(i)} \frac{\partial L}{\partial s_{ij}} \Big|_{\mathbf{S}=\mathbf{S}^*(\mathbf{V}); b=b^*(\mathbf{V})} \frac{\partial s_{ij}^*(\mathbf{V})}{\partial V_{ac}} \\ &+ \frac{\partial L}{\partial b} \Big|_{\mathbf{S}=\mathbf{S}^*(\mathbf{V}); b=b^*(\mathbf{V})} \frac{\partial b^*(\mathbf{V})}{\partial V_{ac}} + \frac{\partial L(\mathbf{S}, \mathbf{V}, b)}{\partial V_{ac}} \Big|_{\mathbf{S}=\mathbf{S}^*(\mathbf{V}); b=b^*(\mathbf{V})}. \end{aligned} \quad (\text{A.14})$$

The double sum on the right vanishes because, for each (i, j) pair, $(\partial L/\partial s_{ij})$ will vanish if s_{ij}^* is interior to $[s_{ij}^-, s_{ij}^+]$, while $(\partial s_{ij}^*/\partial V_{ac})$ vanishes if s_{ij}^* is at a boundary point.

The term $\partial L/\partial b = B(\mathbf{S}^*(\mathbf{V}), \mathbf{V})$ will vanish at optimality due to equation (A.13). This leads to equation (A.8) and eventually to equation (14).

Our reformulation in equations (A.5), (A.6), and (A.2) will not lead to a duality gap because one can always find a unique value of b , since $B(\mathbf{S}^*(\mathbf{V}), \mathbf{V})$ varies smoothly and monotonically with b .

Proof of theorem 2

The throughput constraint (A.3) can be written as

$$1 = Q(TH^+, \mathbf{V}) \equiv \frac{TH^+}{K} \left\{ \sum_{i=1}^M \sum_{j=1}^{n(i)} V_{ij}(s_{ij}^- + D_{ij}) + \left[\frac{K-1}{K} TH^+ \right] \sum_{i \in WC_{FCFS}} V_i \frac{\sum_{j=1}^{n(i)} V_{ij}(s_{ij}^-)^2}{1 - \frac{K-1}{K} TH^+ \sum_{j=1}^{n(i)} V_{ij} s_{ij}^-} \right\}. \quad (A.15)$$

Differentiating both sides of equation (A.15) with respect to V_{ab} , we obtain

$$\frac{\partial Q}{\partial TH^+} \frac{\partial TH^+}{\partial V_{ab}} + \frac{\partial Q}{\partial V_{ab}} = 0, \quad (A.16)$$

which leads to equation (16).

Using equation (A.15), we get

$$\frac{Q(TH^+, \mathbf{V})}{\partial V_{ab}} = \frac{TH^+}{K} (s_{ab}^- + D_{ab}) + \frac{(K-1)(TH^+)^2}{K^2} \frac{\partial(V_a T_a)}{\partial V_{ab}}, \quad (A.17)$$

where

$$T_i = \frac{\sum_{j=1}^{n(i)} V_{ij}(s_{ij}^-)^2}{1 - \frac{K-1}{K} TH^+ \sum_{j=1}^{n(i)} V_{ij} s_{ij}^-}. \quad (A.18)$$

Using the relationship

$$\frac{\partial T_a}{\partial V_{ab}} = \frac{s_{ab}^-}{1 - \frac{K-1}{K} TH^+ + \sum_{j=1}^{n(a)} V_{aj} s_{aj}^-} \left[s_{ab}^- + T_a \frac{K-1}{K} TH^+ \right], \quad (A.19)$$

we get the desired value (equation (18)) for $\partial Q(TH^+, \mathbf{V})/\partial V_{ab}$.

Similarly, we derive equation (17) using equation (A.15):

$$\frac{\partial Q(TH^+, \mathbf{V})}{\partial TH^+} = \frac{1}{K} \sum_{i=1}^M \sum_{j=1}^{n(i)} V_{ij}(s_{ij}^- + D_{ij}) + \frac{K-1}{K} \frac{\partial}{\partial TH^+} \left[\sum_{i \in WC_{FCFS}} V_i \frac{TH^+ \sum_{j=1}^{n(i)} V_{ij}(s_{ij}^-)^2}{1 - \frac{K-1}{K} TH^+ \sum_{j=1}^{n(i)} V_{ij} s_{ij}^-} \right]. \quad (A.20)$$

A simplification of the right-hand side is possible using equation (A.15). This leads to equation (17.)

Appendix B. Proof of theorem 3

Consider the convex knapsack problem:

$$\min \left\{ \sum_{k=1}^{M \rightarrow T} COSTMIN(TH_k) \right\} \quad (B.1)$$

subject to

$$\sum_{k=1}^T TH_k = \sum_{k=1}^T D_k - I_0, \quad (B.2)$$

with the implicit constraints

$$TH^- \leq TH_k \leq TH^+, \quad 1 \leq k \leq T \quad (B.3)$$

and

$$I_0 + \sum_{j=1}^k (TH_j - D_j) \geq 0, \quad 1 \leq k \leq T. \quad (B.4)$$

Define the Lagrangian multiplier

$$L \equiv \frac{\partial COSTMIN(\bar{D})}{\partial TH}. \quad (B.5)$$

The solution pair $\{TH_k = \bar{D} \text{ for } 1 \leq k \leq T, L\}$ is optimal for the convex knapsack problem because it meets the Kuhn-Tucker optimality conditions. Since the single-period costs are *increasing* with TH_k , there is no incentive to produce more than \bar{D} . Using expressions (B.1)–(B.2) as a *relaxation* of the original problem (25)–(28), the solution $\{TH_i = \bar{D}\}$ is feasible for the original problem, and no lower value of the objective function (25) is possible when every $h_i = 0$.

Appendix C. Minimum-cost cutting-speed and cost-function derivations

This appendix presents the derivations of the minimum-cost cutting-speed, time, and cost function for stand-alone machining operations.

C.1. The minimum-cost cutting-speed

It is commonly assumed (Boucher 1987; Drozda and Wicks 1983; Primrose and Leonard 1986) that the variable machining cost per part c_T is approximately

$$c_T = c_d(s + t_l) + c_t \left(\frac{s}{T} \right), \quad (\text{C.1})$$

where

c_d = direct machine and labor operating costs in dollars per minute (\$/minute)

s = machining time (minutes)

t_l = machine idle time per piece (including part loading and unloading)

c_t = tool replacement cost including purchasing and setup cost (\$/piece)

T = tool life (minutes)

The typical relationship between the mean tool life T and the cutting speed v in meters per minute (m/minute) is given by Taylor (1907) and Hitomi (1979):

$$vT^n = c, \quad (\text{C.2})$$

where n ($0 < n < 1$) and c ($c > 0$) are constants. These constants depend upon the tool, workpiece, and machining conditions.

Inserting T from equation (C.2) into equation (C.1), recognizing $s = \text{constant}/v$, and differentiating c_T with respect to v results in v^* , the minimum-cost cutting-speed (for performing this one operation):

$$v^* = c \left[\frac{c_d}{\left(\frac{1-n}{n} \right) c_t} \right]^n. \quad (\text{C.3})$$

The actual value of s in equation (C.1) depends on the part geometry and process involved. For example, in a turning operation, the following relationships hold between the cutting speed, feed rate, and the process time:

$$s = \frac{\pi d L}{1000 v f}, \quad (\text{C.4})$$

where

s = process time in minutes

d = diameter of the workpiece in millimeter (mm)

L = length of workpiece (mm)

v = cutting speed (m/minute)

f = feed rate in mm per revolution (mm/rev)

Studies by Cook (1973), Ramalingam and Watson (1978), Wagner and Barash (1971), and others point to the stochastic aspects of tool wear and tool life and to its impact on manufacturing productivity. However, the analyses by Chang et al. (1982), Fenton and Joseph (1979), and Sheikh et al. (1980) indicate that most deterministic (or mean value models) are fairly close to their stochastic counterparts when prescribing the minimum cost-cutting speeds. This robustness of the deterministic models is explained by the convex structure of the machining cost functions.

C.2. The tool-cost function

The optimization framework given by equations (4)–(12) attempts to minimize the cost per hour for a given throughput level TH . The key observation is that the direct cost per hour of machines and labor is essentially independent of tool-speed choice and throughput choice over a very wide range. The direct cost term—the first term in equation (C.1)—is unaffected by the optimization framework (4)–(12), and is deleted. The tool-cost function to be used here reduces to

$$g(s) = c_t \frac{s}{T}. \quad (C.5)$$

In the case of turning operations, equations (C.2) and (C.4) are inserted into equation (C.5), which then leads to

$$g(s) = c_t \left(\frac{1000cf}{\pi dL} \right)^{-(1/n)} s^{-(1-n)/n}. \quad (C.6)$$

References

- Ayres, R.V., "Future Trends in Factory Automation," *Manufacturing Review*, Vol. 1, No. 2, pp. 93–103 (1988).
- Bitran, G.R. and Tirupati, D., "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 Engineering Economist*, Vol. 32, No. 3, pp. 217–230 (1987).
- Boxma, O.J., Rinnooy Kan, A.H.G., and van Vliet, M., "Machine Allocation Problems in Manufacturing Networks," *European Journal of Operational Research*, Vol. 45, No. 1, pp. 47–54 (1990).
- Buzacott, J.A. and Yao, D.D., "Flexible Manufacturing Systems: a Review of Analytical Models," *Management Science*, Vol. 32, No. 7, pp. 890–907 (1986).
- Chang, T., Wysk, R.A., Davis, R.P., and Choi, B., "Milling Parameter Optimization Through a Discrete Variable Transformation," *International Journal of Production Research*, Vol. 20, No. 4, pp. 507–516 (1982).
- Cook, N.H., "Tool Wear and Tool Life," *ASME, Journal of Engineering for Industry*, Vol. 95, No. 11, pp. 931–938 (1973).
- Cummings, S., "Developing Integrated Tooling Systems: a Case Study at Garret Turbine Engine Company," Proceedings of the Fall Industrial Engineering Conference, Boston, MA (1986).
- Dallery, Y. and Frein, Y., "An Efficient Method to Determine the Optimal Configuration of a Flexible Manufacturing System," in *Proceedings of the Second ORSA/TIMS Conference on Flexible Manufacturing Systems*, K.E. Stecke and R. Suri (Eds.), Elsevier Science Publishers, B.V., Amsterdam (1986).

- Drozda, T.J. and Wick, C., *Tool and Manufacturing Engineering Handbook*, Vol. I, Dearborn, MI, Society of Manufacturing Engineers (1983).
- Eilon, S., "On Smoothing Shipments—A Comment," *Management Science*, Vol. 17, pp. 608–609 (1971).
- Fenton, R.G. and Joseph, N.D., "The Effects of the Statistical Nature of Tool-Life on the Economics of Machining," *International Journal of Machine Tool Design Research*, Vol. 19, pp. 43–61 (1979).
- Gray, A.E., Seidmann, A., and Stecke, K.E., "A Synthesis of Tool Management Issues and Decision Problems in Automated Manufacturing," Working Paper CMOM 88-03 (revised), William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, NY (1990).
- Hax, A.C. and Candea, D., *Production and Inventory Management*, Prentice-Hall, Engelwood Cliffs, NJ (1984).
- Himmelblau, D.M., *Applied Nonlinear Programming*, McGraw-Hill, New York (1972).
- Hitomi, K., "Optimization of Multistage Machining System: Analysis of Optimal Machining Conditions for the Flow-Type Machining System," *ASME Journal Engineering for Industry*, Vol. 73, No. 2, pp. 598–606 (1971).
- Hitomi, K., *Manufacturing Systems Engineering*, Taylor & Francis, London (1979).
- Johnson, L.A. and Montgomery, D.C., *Operations Research in Production Planning, Scheduling and Inventory Control*, Wiley, New York (1974).
- McCartney, J. and Hinds, B.K., "Tooling Economics in Integrated Manufacturing Systems," *International Journal of Production Research*, Vol. 20, No. 4, pp. 493–505 (1982).
- McLeavey, D.W. and Narashiman, S.L., *Production and Inventory Control*, Allyn and Bacon, Boston, MA (1985).
- Nahmias, S., *Production and Operations Analysis*, Irwin, Homewood, IL (1989).
- Olberg, E.F., Jones, D., and Horton, H.L., *Machinery's Handbook*, 20th Edition, Industrial Press, New York (1976).
- Peterson, R., "Optimal Smoothing of Shipments in Response to Orders," *Management Science*, Vol. 17, pp. 597–607 (1971).
- Primrose, P.L. and Leonard, R., "Reappraising Cutting Tool Economics Within the Bounds of Accountancy Theory," *International Journal of Production Research*, Vol. 24, No. 2, pp. 269–278 (1986).
- Ramalingam, S. and Watson, J.D., "Tool Life Distributions. Part 4. Minor Phases in Work Material and Multiple-Injury Tool Failure," *ASME, Journal of Engineering for Industry*, Vol. 100, No. 1, pp. 201–209 (1978).
- Schweitzer, P.J., "Approximate Analysis of Multiclass Closed Networks of Queues," *Proceedings of the International Conference on Stochastic Control and Optimization*, Free University, Amsterdam, Netherlands, pp. 5–16 (1979).
- Schweitzer, P.J. and Seidmann, A., "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).
- Schweitzer, P.J. and Seidmann, A., "Performance Optimization and Capacity Range Analysis for FMSs with Distinct Multiple Job Visits to Work Centers," Working Paper QM89-10, William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, NY (1989) (revised 1990).
- Schweitzer, P.J. and Seidmann, A., "Optimizing Processing Rates for Flexible Manufacturing Systems," *Management Science*, Vol. 37, No. 4, pp. 454–466 (1991).
- Shanthikumar, J.G. and Yao, D.D., "Optimal Server Allocation in a System of Multiserver Stations," *Management Science*, Vol. 33, No. 9, pp. 1173–1180 (1987).
- Shanthikumar, J.G. and Yao, D.D., "On Server Allocation in Multiple Center Manufacturing Systems," *Operations Research*, Vol. 36, No. 2, pp. 333–342 (1988).
- Shanthikumar, J.G. and Yao, D.D., "Second-order Properties of the Throughput of a Closed Queueing Network," *Mathematics of Operations Research*, Vol. 13, pp. 524–535 (1988b).
- Shanthikumar, J.G. and Yao, D.D., "Optimal Allocation of Buffers in a System of Flexible Manufacturing Cells," *International Journal of Flexible Manufacturing Systems*, Vol. 1, No. 4, pp. 347–356 (1989).
- Sheikh, A.K., Kendal, L.A., and Pandit, S.M., "Probabilistic Optimization of Multitool Machinery Operations," *ASME, Journal of Engineering for Industry*, Vol. 102, pp. 239–246 (1980).
- Silver, E.A., "A Tutorial on Production Smoothing and Work Force Balancing," *Operations Research*, Vol. 15, pp. 985–1010 (1967).
- Solot, P., "A Heuristic Method to Determine the Number of Pallets in a Flexible Manufacturing System with Several Pallet Types," *International Journal of Flexible Manufacturing Systems*, Vol. 2, No. 3, pp. 191–216 (1990).
- Solot, P. and Bastos, J.M., "MULTIQ: A Queueing Model for FMSs with Several Pallet Types," *Journal of the Operational Research Society*, Vol. 39, No. 9, pp. 811–821 (1988).

- Stecke, K.E., "Algorithms for Efficient Planning and Operation of a Particular FMS," *International Journal of Flexible Manufacturing Systems*, Vol. 1, No. 4, pp. 287-324 (1989).
- Suri, R., "RMT Puts Manufacturing at the Helm," *Manufacturing Engineering*, Vol. 100, pp. 41-44 (February 1988).
- Taylor, F.W., "On the Art of Cutting Metals," *Transactions of the American Society of Mechanical Engineers*, Vol. 28, pp. 31-350 (1907).
- Trucks, H.E., *Designing for Economical Production*, Society of Manufacturing Engineers, Dearborn, MI (1987).
- Vinod, B. and Solberg, J.J., "The Optimal Design of Flexible Manufacturing Systems," *International Journal of Production Research*, Vol. 23, No. 6, pp. 1141-1151 (1985).
- von Vliet, M.V. and Rinnooy Kan, A.H.G., "Machine Allocation Algorithms for Job Shop Manufacturing," to appear in *Journal of Intelligent Manufacturing* (1990).
- Wagner, J.G. and Barash, M.M., "Study of the Distribution of the Life of HSS Tools," *ASME, Journal of Engineering for Industry*, Vol. 93, No. 11, pp. 1044-1050 (1971).
- Watanabe, T. and Fujii, R., "Determining the Job Operation Speed and Schedule for Machine Tools in FMS by Forecasting Using a Simulator and a System Performance Index," *ASME Proceedings of the USA-Japan Symposium on Flexible Automation*, Minneapolis, MN, pp. 751-756 (1988).