

Dynamic Load Control Policies for a Flexible Manufacturing System with Stochastic Processing Rates

ABRAHAM TENENBAUM

Department of Industrial Engineering, Tel Aviv University, Tel Aviv, 69978, Israel

ABRAHAM SEIDMANN

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

Abstract. Real-time scheduling and load controls of FMSs are complex processes in which the control logic must consider a broad spectrum of instantaneous state variables while taking into account the probabilistic future impact of each decision at each time epoch. These processes are particularly important in the management of modern FMS environment, since they are known to have a significant impact on the FMS productive capacity and economic viability. In this article we outline the approach developed for dynamic load controls within an FMS producing a variety of glass lenses. Two revenue-influencing objective functions are evaluated for this capital-intensive facility. It is shown that by using Semi-Markovian modeling concepts, the FMS states need to be observed only at certain decision epochs. The mean holding time in each state is then obtained using the probability distribution function of the conditional state occupancy times. Several key performance measures are then derived by means of the value equations. In addition, the structure of the optimal policies are exemplified for a variety of operational parameters. It is shown that the optimal policies tend to generate higher buffer stocks of parts in those work centers having the highest revenue-generation rates. These buffer stocks get smaller and smaller as the relative processing capacity of the centers increases. Similar observations lead us to the introduction of several promising heuristics that capture the structural properties of the optimal policies with a significantly smaller computational effort. Results of the empirical evaluation of these heuristics are also analyzed here.

1. Background

This article deals with developing efficient approaches for Dynamic Load Controls (DLCs) for certain Flexible Manufacturing Systems (FMSs). An FMS can be defined as a set of processing stations on which a variety of parts are transformed and between which these parts are automatically transported under the real-time control of a computer system. Recent surveys of current practice and research being done on FMSs can be found in Drozda (1988), Gershwin et al. (1986), Gray et al. (1988), Kusiak (1986), and in Stecke and Suri (1986).

The DLC scheme presented here was motivated by planning a new FMS in an electro-optical manufacturing plant. This FMS was designed to produce a variety of high-precision glass lenses in small batches and with minimal setup times. In this facility, a special metallic alloy, having a low melting point, is used to bond each raw lens to a material handling pallet. These pallets, with their attached lenses, go through two processing stages in the FMS: coarse and fine grinding in the first, and precision honing and polishing in the second. Several parallel and identical Direct Numerical Control (DNC) machining centers with automatic tool changers compose the first processing stage. Each center operates its

own tool magazine, which enables storage and fast retrieval of the tools required to perform all the operations on a family of parts (lenses). The use of large-capacity tool magazines and of Automated Tool Changers (ATCs) eliminated the setup times required to change from one part type to another. The basic premise for this flexibility is to allow for better load balancing and for faster responses to the concurrent needs of the second-stage stations. The second processing stage is made up of another set of several dedicated *workstations*, where each is designed to fabricate a special lens type. A substantial saving in production time is obtained by using two dedicated programmable manipulators in each machining center and workstation. Each pair of these manipulators simultaneously unloads the pallet with the processed lens and loads the pallet with the next one. Figure 1 outlines the physical layout of the system. A fast rack-and-pinion transporter carries the pallets from the machining centers to the workstations. The expected utilization of that transporter is below 30% in this layout, and the processing rate is significantly smaller than the transportation rate. Since the transporter hardly affects the FMS performance, it was decided to exclude it from the current DLC model. For brevity, the machining centers will be denoted, in this paper, as *centers* and the dedicated workstations, operating in the second stage, as *stations*.

Limited input buffer capacity is available at each station. Such buffers are introduced to improve the overall performance of the FMS by decoupling the cells and the stations (Okamura and Yamashina 1979; Gershwin and Berman 1981). The system was designed with small buffers in order to control the WIP (work in progress) levels, to reduce hardware costs, and to provide immediate quality feedback to the preceding processing stage. Instantaneous yield fluctuations in the centers and stations require the application of a rigorous DLC policy to meet the production requirements while minimizing the blocking effects. Hence, whenever a center completes a lens, the system controller has to determine the part (lens) type to process next. The DLC problem in the FMS described above is, therefore, the selection of the part type to be processed next at each center as a function of the current status of all other cells and of the buffer stocks in each station.

Recent studies of similar systems with limited buffers include Alam et al. (1985), Buzacott (1985), Elsayed and Hwang (1986), Hahne (1981), Seidmann and Tenenbaum (1989), Suri and Diehl (1986) and Shanthikumar and Yao (1989). Several other studies of FMS scheduling with stochastic processing rates include Akella et al. (1985), Hildebrant (1980), and Shalev-Oren et al. (1985). Most of these authors point to the potential improvements in the manufacturing system performance that are attainable through the proper use of dynamic control strategies based on a rational function of the current queue lengths in the local buffers. However, these studies assume simple parts flows and do not treat the (more realistic) case of multiple part types flow through multistage parallel heterogeneous machines.

The major objective of the analysis reported here is the explicit derivation of tractable optimal and heuristic DLC policies. Two major criteria are considered. The first one aims at minimizing the weighted starvation penalty incurred by the stations when they run out of input parts; these penalties are surrogates for idle time and lost production opportunities. The second criteria aims at maximizing the expected throughput rate of the system. Both criteria are revenue-influencing surrogates and are commonly used in capital-intensive facilities (Rachamadugu and Stecke 1989, Smith et al. (1986). Other studies dealing with throughput rate maximization for FMSs operating in a stochastic environment include Han and McGinnis (1988), Masri and Hausman (1988), Pinedo et al. (1986), Seidmann and Schweitzer (1984) and Stecke (1989).

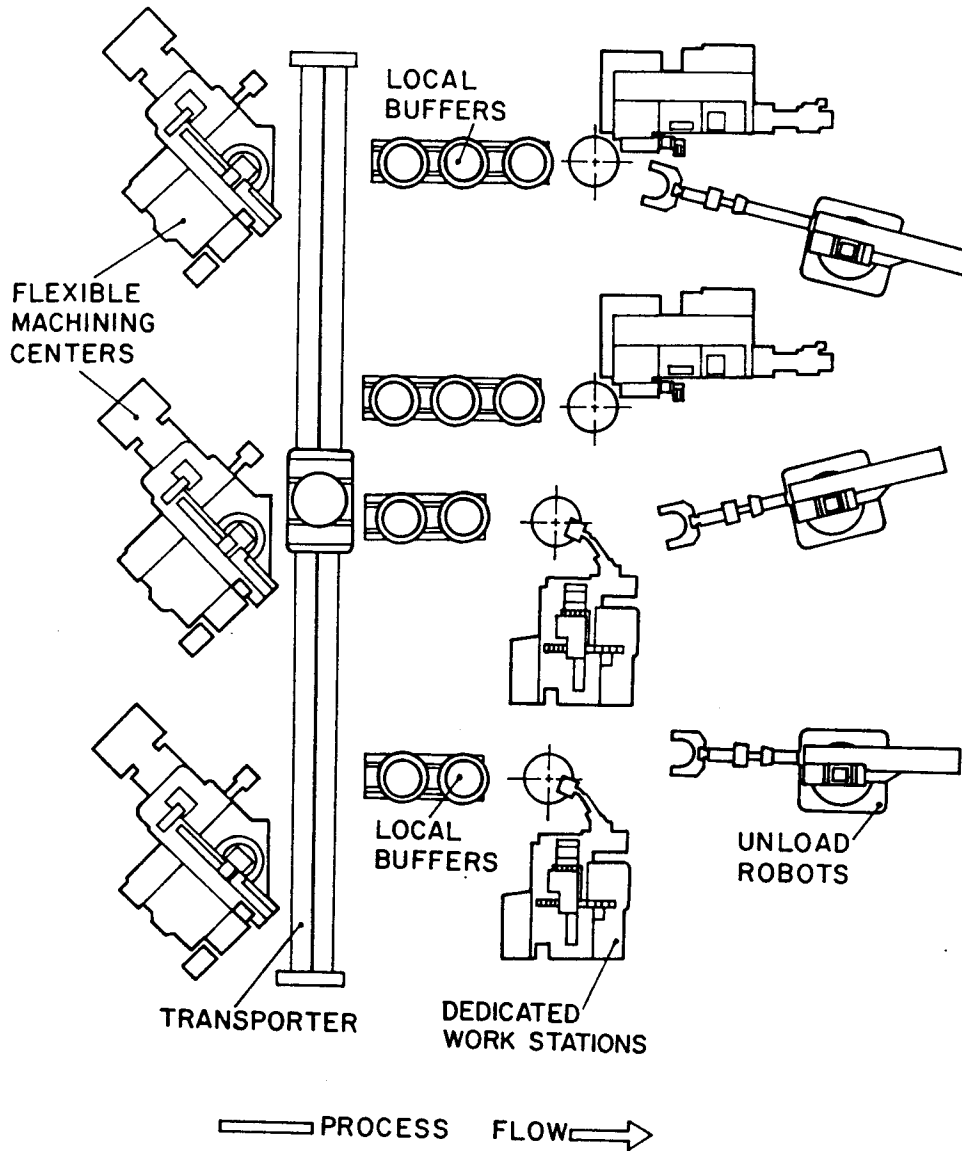


Figure 1. Schematic layout of the flexible glass lens manufacturing facility. This chart shows the three centers ($S = 3$) and the four ($R = 4$) stations. The buffers capacities (B_i) are four and three for the two top and the two bottom stations, respectively (i.e., $B_1 = B_2 = 4$, $B_3 = B_4 = 3$).

The FMS basic model is described and its functional equations are formulated in Section 2. Several heuristic policies are developed in Section 3 and are illustrated later in Section 4. Section 5 concludes this article, and the Appendix formulates the exact derivation of the FMS performance measures for any given policy.

2. The FMS modeling approach

2.1. Introduction

This section presents the major components of the FMS model, the decision-epochs state space and the feasible set of transitions between these states; the mathematical formulation of the stochastic dynamic programming functional equations and their finite-differences solution are also described here for the two objective functions defined above.

2.2. Parameters and variables

The model assumes that the FMS consists of $S(\geq 1)$ parallel and identical centers producing parts for $R(\geq 1)$ stations. Each part is first processed at one of the centers and then routed for additional processing at one of the stations. There is always a supply of pallets with raw parts available to the centers, and the processing time for each part at the center depends on its designated downstream station. The time required for a center to process one part for station i (part type i) is exponentially distributed with a mean of $1/\mu_i$, $1 \leq i \leq R$. Each station processes a certain type of parts and has its own buffer and production capabilities. It has processing time that is exponentially distributed with a mean of $1/\lambda_i$ ($1 \leq i \leq R$) and a finite local input buffer with room for $B_i (\geq 1)$ parts, one in processing plus up to $B_i - 1$ on queue. The exponential distribution assumption allows us to gain insights into the dynamic behavior of the FMS even if this assumption is not quite the case in practice (Yao and Buzacott 1986). Transport times between the centers and the workstations are relatively small, and ignoring them has negligible effects on the model's accuracy. Parts leave the FMS after being processed at the stations. Station i , when idle because of the absence of parts, incurs a shortage (starvation) penalty of C_i ($1 \leq i \leq R$) dollars per hour idle. Alternatively, we consider a weight or a contribution of w_i dollars ($w_i \geq 0$, $1 \leq i \leq R$) for each item produced at station i . Production control decisions are made before each part is processed by a center and are made with full knowledge of the instantaneous processing status of the centers, and of the buffer status at all the R stations. Figure 2 describes the schematic flow of parts, status data, and decisions in this facility.

The proposed methodology is expected to be easily implemented in the FMS control software. Following the initial computation of the optimal policy, the production controller is constantly fed with the discrete changes in the FMS state. The control policies discussed below generate a set of precalculated lookup tables (flat files) prescribing the desired response for each system state. Using this set of precalculated lookup tables, the FMS controller retrieves the desired discrete response and then assigns the various part types to the appropriate centers. Having several such tables, the control policy easily responds to changes in the availability of centers and stations (due to setups and corrective or preventive maintenance), to new part routes, or to changes in the required processing times. The relative starvation penalties and the marginal profit contribution of each part type are evaluated by a higher level of the control hierarchy that encompasses such issues as parts-mix changes and plantwide workload balancing (Suri and Whitney 1984). Each time the penalty and profit parameters are varied, one needs to recompute and reload these lookup tables.

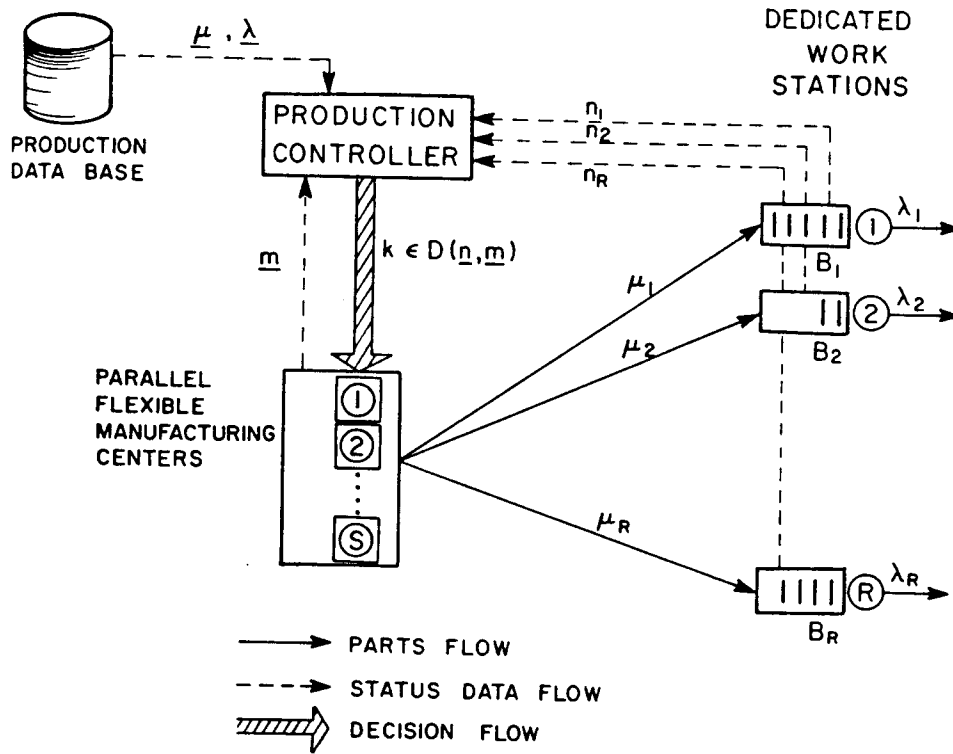


Figure 2. Parts and control information flows within the FMS. Note that n_i is the number of items on buffer i , $1 \leq i \leq R$, and m_j , $1 \leq j \leq S$, is the number of centers currently processing parts for station j .

2.3. State space and admissible decision sets

We let $\Omega = \{ \mathbf{n} = (n_1, n_2, \dots, n_R) \mid 0 \leq n_i \leq B_i \}$ be the state space of the buffers, where n_i ($i = 1, 2, \dots, R$) denotes the number of parts in station i . When all the buffers are full (blocked) then $\mathbf{n} = \mathbf{B} \triangleq (B_1, B_2, \dots, B_R)$. The number of centers currently producing parts for station ($i = 1, 2, \dots, R$) is defined as m_i . The centers can be either blocked or unblocked, depending on the state of the buffers. No cell can remain idle if there exists an empty buffer that expects a new part. Decisions are made when at least one center is idle and there is also at least one empty space at the buffers. In addition, the total number of parts produced for station i can exceed neither the number of centers (S), nor the instantaneous number of vacant buffer spaces at that station ($B_i - n_i$). Hence a decision epoch can be reached only following the *end-of-processing event* at one (or more) of the centers. This leads to the following definition of Φ to be used later in characterizing the state of the centers at the decision epochs:

$$\Phi = \{ \mathbf{m} (m_1, \dots, m_R) \mid m_i = 0, \dots, \text{Min}(S - 1, B_i - n_i), i = 1, \dots, R \}.$$

For brevity, we define here

$$| \mathbf{m} | = \sum_{i=1}^R m_i, \quad (1)$$

$$| \mathbf{B} | = \sum_{i=1}^R B_i, \quad (2)$$

$$| \mathbf{n} | = \sum_{i=1}^R n_i. \quad (3)$$

The complete state space of decision epochs in this system is a function of both \mathbf{n} (the buffers' status) and \mathbf{m} (the stations' status). It is next described as the union of the following four mutually exclusive decision regions (several numerical illustrations of these decision regions are provided by Example 1 in Section 4).

2.3.1. The initial state.

$$| \mathbf{m} | = 0 \text{ and } | \mathbf{n} | = 0, \quad S > 1.$$

In this initial state the buffers are empty, the stations idle, and the S centers are scheduled to start processing immediately. It means that the controller has to make S decisions defined by the optimal decision vector \mathbf{M} ($\mathbf{M} \triangleq M_1, M_2, \dots, M_R$), where M_i ($i = 1, 2, \dots, R$) is the number of centers scheduled for processing type i parts. Activating all S centers means that the admissible decision space for \mathbf{M} is

$$\delta = \{ \mathbf{M} = (M_1, M_2, \dots, M_R) \mid 0 \leq M_i \leq \text{Min}(S, B_i) \text{ and } | \mathbf{M} | = S, (i = 1, 2, \dots, R) \}.$$

Since all the cells will be occupied, we also get

$$| \mathbf{M} | = \sum_{i=1}^S M_i = S. \quad (4)$$

The mean time that the system will hold in this state is

$$1/\mu(\mathbf{M}) \quad \text{where } \mu(\mathbf{M}) = \sum_{i=1}^R M_i \mu_i. \quad (5)$$

The next decision epoch will be at an end-of-processing event at one of the centers. The probability that this epoch will be at a center processing a type i part is $\mu_i M_i / \mu(\mathbf{M})$. This condition is illustrated by state number 1 in Tables 1 and 2. Note that this decision region, as well as the third decision region described below, holds only for an FMS having $S > 1$. In fact, the state description of the initial state for $S = 1$ is imbedded in the next region.

Table 1. Optimal processing decisions and the relative state values for an FMS having five work cells: two centers and three stations

State number SN	State n	Space m	Optimal decision M*, k(*)	State value v(n, m)	State number SN	State n	Space m	Optimal decision M*, k(*)	State value v(n, m)
1	0 0 0	0 0 0	020	0.00	261	3 4 4	0 0 0	100	-139.28
2	0 0 0	0 0 1	010	-1.10	262	3 4 4	1 0 0	000	-139.28
3	0 0 0	0 1 0	010	0.00	263	4 0 0	0 0 1	010	-48.59
4	0 0 0	1 0 0	010	12.83	264	4 0 0	0 1 0	010	-49.70
5	0 0 1	0 0 1	010	-25.76	265	4 0 1	0 0 1	010	-73.34
161	2 2 0	0 1 0	001	-78.70	266	4 0 1	0 1 0	010	-77.67
162	2 2 0	1 0 0	001	-74.06	301	4 4 1	0 0 1	001	-132.27
163	2 2 1	0 0 1	010	-100.33	302	4 4 2	0 0 1	001	-137.06
164	2 2 1	0 1 0	110	-100.33	303	4 4 3	0 0 0	001	-139.30
165	2 2 1	1 0 0	001	-97.17	304	4 4 3	0 0 1	000	-139.30
166	2 2 2	0 0 1	010	-112.28	305	4 4 4	0 0 0	000	-140.66
250	3 3 3	1 0 0	010	-133.39					
251	3 3 4	0 1 0	100	-136.89					
252	3 3 4	1 0 0	010	-136.89					
253	3 4 0	0 0 1	001	-113.17					
254	3 4 0	1 0 0	001	-108.24					
255	3 4 1	0 0 1	001	-127.49					

Table 2. Optimal processing decisions and the relative state values for an FMS having seven work cells: four centers and three stations

State number SN	State n	Space m	Optimal decision M*, k(*)	State value v(n, m)	State number SN	State n	Space m	Optimal decision M*, k(*)	State value v(n, m)
1	0 0 0	0 0 0	031	0.00	538	3 3 3	1 1 1	000	-129.99
2	0 0 0	0 0 3	010	5.33	539	3 3 4	0 1 0	100	-132.95
3	0 0 0	0 1 2	010	1.42	540	3 3 4	1 0 0	010	-132.95
4	0 0 0	0 2 1	010	0.00	541	3 3 0	1 1 0	000	-132.95
5	0 0 0	0 3 0	001	0.00	542	3 4 0	0 0 3	001	-111.36
389	2 2 0	2 0 1	001	-69.80	543	3 4 0	1 0 2	001	-109.18
390	2 2 0	2 1 0	001	-68.30	551	3 4 3	1 0 1	000	-133.97
391	2 2 1	0 0 3	010	-98.67	552	3 4 4	0 0 0	100	-136.20
392	2 2 1	0 1 2	010	-99.21	553	3 4 4	1 0 0	000	-136.20
393	2 2 1	0 2 1	001	-99.21	554	4 0 0	0 0 3	010	-42.71
394	2 2 1	1 0 2	010	-97.32	555	4 0 0	1 0 2	001	-46.93
395	2 2 1	1 1 1	010	-97.41	611	4 4 2	0 0 1	001	-133.60
396	2 2 1	1 2 0	001	-97.41	612	4 4 2	0 0 2	000	-133.60
397	2 2 1	2 0 1	010	-94.23	613	4 4 3	0 0 0	001	-136.26
398	2 2 1	2 1 0	001	-94.23	614	4 4 3	0 0 1	000	-136.26
399	2 2 2	0 1 2	010	-110.28	615	4 4 4	0 0 0	000	-138.07

2.3.2. Full centers operation. The full centers operations state refers to a case in which $(S - 1)$ centers are active and one center, which has just completed processing a certain part, is waiting for a new decision:

$$\begin{array}{lll} |m| = S - 1, & 0 \leq |n| \leq |B| - S, & S \geq 1. \\ \text{(one idle center)} & \text{(there is at least one empty input buffer)} & \end{array}$$

The part type scheduled for processing by this center is selected from the following admissible decision space:

$$D(n, m) = \{k \mid n_k + m_k < B_k\}. \quad (6)$$

This definition of $D(n, m)$ ensures that all centers processing parts type k will find empty buffers at station k . This requirement means that for each station (i), the total number of parts at the input buffer (n_i) and in processing by the centers (m_i), will never exceed its maximum buffer capacity (B_i). For example, suppose that station i has three empty buffers ($B_i - n_i = 3$) and currently there is one center ($m_i = 1$) processing parts designated for that station. Since $B_i - n_i = 3 > m_i = 1$, one can safely assign an additional center for processing a part type i without later being blocked. This condition is also illustrated by state number 395 in Table 2.

The mean time that the system will hold in this state, following a decision $k \in D(n, m)$, is $1/\mu(m, k)$, where

$$\mu(m, k) = \sum_{i=1}^R (m_i \mu_i) + \mu_k, \quad (7)$$

and the next decision epoch will be at an end-of-processing event at one of the *centers*.

2.3.3. Partial centers operation with partial blocking. At these decision epochs, the number of idle centers is greater than one. These centers are not all active, since at the previous state the number of active centers (m) was equal to the total number of buffer spaces (i.e., $|m| = |B| - |n|$). This imposes the following ranges for n and m :

$$0 \leq |m| < S - 1, \quad |m| + |n| = |B| - 1, \quad S > 1.$$

Note that the $|B| - 1$ term above results from reaching the current state through an end-of-processing event at one of the *stations*. For example, suppose that the total number of buffers is 12 ($= |B|$). If there are already 10 ($= |n|$) items at the stations and one center is also currently busy ($|m| = 1$), then only one additional center can be assigned a part for production without violating the following policy constraint: $|n| + |m| \leq |B|$. This condition is clearly illustrated by state number 540 in Table 2.

Decrementing $|n|$ allows for scheduling an additional idle center. The part type scheduled for processing by this center is selected from $D(n, m)$ as defined by Equation (6). The

mean time that the system will hold in this state following decision $k \in D(\mathbf{n}, \mathbf{m})$ is $1/(\mu(\mathbf{m}, k) + \lambda(\mathbf{n}))$, where $\lambda(\mathbf{n})$ is the *cumulative processing rate* of all active stations at buffer state (\mathbf{n}) . This rate is defined by

$$\lambda(\mathbf{n}) = \sum_{i \in A(\mathbf{n})} \lambda_i, \quad (8)$$

where $A(\mathbf{n})$ is the set of all *active* stations:

$$A(\mathbf{n}) = \{ i \mid 0 < n_i \leq B_i, 1 \leq i \leq R \}. \quad (9)$$

The next decision epoch will be at an end-of-processing event at either one of the stations or at one of the centers.

The probability that the *next* decision epoch will be at an end-of-processing event at station i ($i = 1, 2, \dots, R$) is

$$\frac{\lambda_i}{\lambda(\mathbf{n}) + \mu(\mathbf{m}, k)},$$

and the probability that this decision epoch will be at an end-of-processing event at a center processing part type i ($i = 1, 2, \dots, R$) is

$$\frac{\mu(m_i, k)}{\lambda(\mathbf{n}) + \mu(\mathbf{m}, k)},$$

where

$$\mu(m_i, k) = \begin{cases} (m_i + 1)\mu_i & \text{if } i = k \\ m_i\mu_i & \text{otherwise.} \end{cases} \quad (10)$$

If the *next* end-of-processing event is at one of the stations, then $|\mathbf{n}|$ is *decremented*, and an additional part can start processing at one of the idle centers. If, on the other hand, the next event is at one of the centers, then $|\mathbf{n}|$ is *incremented*, and that cell remains idle since $|\mathbf{m}| + |\mathbf{n}|$ becomes equal to $|\mathbf{B}|$ (this means that the FMS has just reached a complete blocking state). These complete blocking states are formalized next.

2.3.4. Partial centers operation with complete blocking. The total number of parts in the buffers $|\mathbf{n}|$ and in processing by the centers $|\mathbf{m}|$ is now equal to the total available buffer space $|\mathbf{B}|$. Inactive centers must remain idle until the end of a processing event at one of the stations will decrement $|\mathbf{n}|$. The feasible range of $|\mathbf{n}|$ and $|\mathbf{m}|$ is now

$$0 \leq |\mathbf{m}| \leq S - 1, \quad |\mathbf{m}| + |\mathbf{n}| = |\mathbf{B}|, \quad S \geq 1.$$

For example, suppose that the total number of buffers is 12 ($= |B|$). If there are already 10 ($= |n|$) items at the stations and two cells are busy ($|m| = 2$), then all additional centers must remain idle until n is decremented. This condition is depicted by states number 551 and 612 in Table 2.

The time that the system will hold in this state is $1/(\mu(m) + \lambda(n))$.

The next decision epoch will be at the end of a processing event at one of the *stations* or *centers*. The probability that the next decision epoch will be at the end of a processing event at *station* i ($i = 1, 2, \dots, R$) is

$$\frac{\lambda_i}{\lambda(n) + \mu(m)},$$

where $\mu(m)$ is the cumulative production rate of the centers:

$$\mu(m) = \sum_{i=1}^R m_i \mu_i. \quad (11)$$

The probability that the next decision epoch will be at the end of a processing event at a *center* processing part type i ($i = 1, 2, \dots, R$) is

$$\frac{m_i \mu_i}{\lambda(n) + \mu(m)}.$$

Following this description of the four decision regions, the state space of the system ($n \cdot m$) can be written in a compact way as

$$\Theta = \left\{ \begin{array}{l} n \in \Omega \\ m \in \Phi \\ n = (n_1, \dots, n_R) \\ m = (m_1, \dots, m_R) \end{array} \left| \begin{array}{l} |n| = 0, |m| = 0 \\ 0 \leq |n| \leq |B| - S, |m| = S - 1 \\ |m| + |n| = |B| - 1, 0 \leq |m| < S - 1 \\ |m| + |n| = |B|, 0 \leq |m| \leq S - 1 \end{array} \right. \begin{array}{l} S > 1; \\ S \geq 1; \\ S > 1; \\ S \geq 1. \end{array} \right\} \quad (12)$$

2.4. The functional equations minimizing the expected shortage penalty costs

Given an FMS as described above, it becomes desirable to obtain a discrete and stationary feedback control strategy that minimizes the expected shortage penalty cost per unit time. This leads to the following dynamic programming functional equations (Jewell 1963; Cinlar 1967; Howard 1971). These equations correspond to the four decision regions described above. In the following equations we let e^k be the unit vector in the k th direction, $1 \leq k \leq R$, and define g as the *long-run* expected shortage penalty cost per unit time following the optimal DLC policy:

$$v(o, o) = \text{Min}_{M \in \delta} \left\{ q_1(o, \mu(M)) - g/\mu(M) + \sum_{i=1}^R [M_i \mu_i / \mu(M)] v(e^i, M - e^i) \right\} \quad (13)$$

$$S > 1, |n| = o, |m| = 0.$$

(Initial State)

In Equation (13), the term $q_1(\mathbf{o}, \mu(\mathbf{M}))$ is the one-transition immediate cost. It is the expected cost incurred during the time interval between the production startup at the centers (when all centers and stations are idle) until the *first* part is completed and transferred to the stations. The mean length of this interval is given by Equation (5). Hence, $g/\mu(\mathbf{M})$ is the time-average cost incurred. At the initial state all the buffers are empty, and therefore

$$q_1(\mathbf{o}, \mu(\mathbf{M})) = \sum_{i=1}^R C_i / \mu(\mathbf{M}). \quad (14)$$

The next set of decisions is derived from

$$\begin{aligned} v(\mathbf{n}, \mathbf{m}) = & \text{Min}_{k \in D(\mathbf{n}, \mathbf{m})} \{q_2(\mathbf{n}, \mu(\mathbf{m}, k)) - g / \mu(\mathbf{m}, k) \\ & + \sum_{i=1}^R [\mu(m_i, k) / \mu(\mathbf{m}, k)] \sum_{\mathbf{o} \leq \mathbf{j} \leq \mathbf{n}} p(\mathbf{j} | \mathbf{n}, \mu(\mathbf{m}, k)) v(\mathbf{n} - \mathbf{j} + \mathbf{e}^i, \mathbf{m} + \mathbf{e}^k - \mathbf{e}^i)\} \quad (15) \\ & \{\mathbf{n}, \mathbf{m}\} \in \Theta, \quad |\mathbf{m}| = S - 1, \quad 0 \leq |\mathbf{n}| \leq |\mathbf{B}| - S, \quad S \geq 1. \\ & \text{(Full Centers Operation)} \end{aligned}$$

The one-transition immediate cost in Equation (15) corresponds to decision k , buffers state \mathbf{n} , and a mean time until the next production decision that is equal to $1/\mu(\mathbf{m}, k)$. This cost is given by

$$q_2(\mathbf{n}, \mu(\mathbf{m}, k)) = \sum_{i=1}^R C_i U(n_i, \mu(\mathbf{m}, k)). \quad (16)$$

We define $U(n_i, \mu(\mathbf{m}, k))$ as the mean period within the time interval before the next decision epoch during which station i is idle (starved), given that this time interval began with n_i parts at station i . It results from the fact that a busy station may become idle at any instant between consecutive decision epochs. This idle time period is clearly equal to

$$U(n_i, \mu(\mathbf{m}, k)) = \frac{1}{\mu(\mathbf{m}, k)} \left[\frac{\lambda_i}{\lambda_i + \mu(\mathbf{m}, k)} \right] n_i \quad (17)$$

$$1 \leq i \leq R, \quad 0 \leq n_i \leq B_i.$$

We next need to characterize the transition probabilities of the system states. Define these by $P(\mathbf{j} | \mathbf{n}, \mu(\mathbf{m}, k))$, with $\mathbf{j} = (j_1, j_2, \dots, j_R)$ the *joint* probability that station i uses up j_i of its n_i ($j_i \leq n_i$) parts during the time interval until the next decision epoch. This joint probability is computed by replacing μ_k by $\mu(\mathbf{m}, k)$ in Equation (15) of Seidmann and Schweitzer (1984). The probability that, following a decision $k \in D(\mathbf{n}, \mathbf{m})$, the next decision epoch will be an end-of-processing event at a center processing part type i is

$$\mu(m_i, k) / \mu(\mathbf{m}, k).$$

Recall that $(\mathbf{n} - \mathbf{j} + \mathbf{e}^i)$ and $(\mathbf{m} + \mathbf{e}^k - \mathbf{e}^i)$ prescribe the state of the buffers and the centers, respectively, at the *next* decision epoch (\mathbf{n} drops by \mathbf{j} , one center adds a part to station i , and decision k was made at state \mathbf{m} of the cells). The expected future value of this state as a function of the decision variable $k \in D(\mathbf{n}, \mathbf{m})$ is given by the summed product of the following terms:

$$\sum_{i=1}^R [\mu(m_i, k) / \mu(\mathbf{m}, k)] \sum_{\mathbf{0} \leq \mathbf{j} \leq \mathbf{n}} P(\mathbf{j} | \mathbf{n}, \mu(\mathbf{m}, k)) v(\mathbf{n} - \mathbf{j} + \mathbf{e}^i, \mathbf{m} + \mathbf{e}^k - \mathbf{e}^i).$$

The final set of decisions is derived from

$$\begin{aligned} v(\mathbf{n}, \mathbf{m}) = & \text{Min}_{k \in D(\mathbf{n}, \mathbf{m})} \{q_3(\mathbf{n}, \mu(\mathbf{m}, k) + \lambda(\mathbf{n})) - g / (\mu(\mathbf{m}, k) + \lambda(\mathbf{n})) \\ & + \sum_{i=1}^R [\mu(m_i, k) / (\mu(\mathbf{m}, k) + \lambda(\mathbf{n}))] v(\mathbf{n} + \mathbf{e}_i, \mathbf{m} + \mathbf{e}^k - \mathbf{e}^i) \\ & + \sum_{i \in A(\mathbf{n})} [\lambda_i / (\mu(\mathbf{m}, k) + \lambda(\mathbf{n}))] v(\mathbf{n} - \mathbf{e}^i, \mathbf{m} + \mathbf{e}^k)\} \end{aligned} \quad (18)$$

$$\{\mathbf{n}, \mathbf{m}\} \in \Theta, S > 1, 0 \leq |\mathbf{m}| < S - 1, |\mathbf{m}| + |\mathbf{n}| = |\mathbf{B}| - 1.$$

(Partial Centers Operations with Partial Blocking)

The one-transition immediate cost in Equation (18) is

$$q_3(\mathbf{n}, \mu(\mathbf{m}, k) + \lambda(\mathbf{n})) = \sum_{i \in E(\mathbf{n})} C_i / (\mu(\mathbf{m}, k) + \lambda(\mathbf{n})), \quad (19)$$

where $E(\mathbf{n})$ is the set of all idle (starving) stations:

$$E(\mathbf{n}) = \{i \in (1, 2, \dots, R) \mid n_i = 0, \quad 1 \leq i \leq R\}.$$

The structure of the immediate cost used in this equation is based on the observation that, following a decision k , the status of both centers ($= \mathbf{m} + \mathbf{e}^k$) and the station ($= \mathbf{n}$) remains *constant* until the next decision epoch. Then, the system will end up either in state $(\mathbf{n} + \mathbf{e}^i, \mathbf{m} - \mathbf{e}^k - \mathbf{e}^i)$ or in state $(\mathbf{n} - \mathbf{e}^i, \mathbf{m} + \mathbf{e}^k)$, depending on whether that next epoch will be an end-of-processing event at one of the centers, or at one of the stations, respectively. The expected future values of these two kinds of decisions epochs are given by the third and fourth terms in Equation (18).

The blocked case requires no decision:

$$\begin{aligned}
 v(\mathbf{n}, \mathbf{m}) &= q_4(\mathbf{n}, \mu(\mathbf{m}) + \lambda(\mathbf{n})) - g/(\mu(\mathbf{m}) + \lambda(\mathbf{n})) \\
 &+ \sum_{i=1}^R [\mu_i m_i / (\mu(\mathbf{m}) + \lambda(\mathbf{n}))] v(\mathbf{n} + \mathbf{e}^i, \mathbf{m} - \mathbf{e}^i) \\
 &+ \sum_{i \in A(\mathbf{n})} [\lambda_i / (\mu(\mathbf{m}) + \lambda(\mathbf{n}))] v(\mathbf{n} - \mathbf{e}^i, \mathbf{m}) \quad (20)
 \end{aligned}$$

$$\{\mathbf{n}, \mathbf{m}\} \in \Theta, S \geq 1, 0 \leq |\mathbf{m}| \leq S - 1, |\mathbf{m}| + |\mathbf{n}| = |\mathbf{B}|.$$

(Partial Centers Operation with Complete Blocking)

In the case of partial cells operations with complete blocking (Equation (20)), the one-transition immediate cost represents the sum of the starvation penalties over all idle stations times the expected time to the next decision epoch. This cost is given by

$$q_4(\mathbf{n}, \mu(\mathbf{m}) + \lambda(\mathbf{n})) = \sum_{i \in E(\mathbf{n})} C_i / (\mu(\mathbf{m}) + \lambda(\mathbf{n})). \quad (21)$$

Since complete blocking is encountered, no decision can be made. It means that the system can evolve either to state $(\mathbf{n} + \mathbf{e}^i, \mathbf{m} - \mathbf{e}^i)$ or to state $(\mathbf{n} - \mathbf{e}^i, \mathbf{m})$, depending on whether that next epoch is the end of a processing event at the centers or at the stations. Equation (22) is used to fix the arbitrary additive component of the relative value vector $v(\mathbf{n}, \mathbf{m})$:

$$v(\mathbf{o}, \mathbf{o}) = 0, \quad \text{(Transient State)} \quad (22)$$

2.5. Maximizing the expected throughput contribution

We now consider the case where a contribution of w_i dollars ($w_i \geq 0, 1 \leq i \leq R$) is associated with each part completed by station i . Generating the optimal policy that maximizes the weighted throughput contributions follows the same line as in Section 2.4. The functional equations to be used are similar to Equations (13), (15), (18), and (20) except for the following changes: the Min operators are replaced by the Max operators and the one-transition costs (now rewards) terms are now

$$q_1(\mathbf{o}, \mu(\mathbf{m})) = \sum_{i=1}^R w_i \frac{n_i \mu_i}{\mu(\mathbf{n})}, \quad (23)$$

$$q_2(\mathbf{n}, \mu(\mathbf{m}, k)) = \sum_{i=1}^R w_i \frac{\mu(m_i, k)}{\mu(\mathbf{m}, k)}, \quad (24)$$

$$q_3(\mathbf{n}, \mu(\mathbf{m}, k) + \lambda(\mathbf{n})) = \sum_{i=1}^R w_i \frac{\mu(m_i, k)}{\mu(\mathbf{m}, k) + \lambda(\mathbf{n})}, \quad (25)$$

$$q_4(\mathbf{n}, \mu(\mathbf{m}) + \lambda(\mathbf{n})) = \begin{cases} \sum_{i=1}^R w_i \frac{\mu_i m_i}{\mu(\mathbf{m}) + \lambda(\mathbf{n})} & \text{if } \mathbf{n} \neq \mathbf{0} \\ 0 & \text{otherwise.} \end{cases} \quad (26)$$

The computed value of g in the modified model reveals the expected long-run contribution per unit time following the optimal production policy.

2.6. Computational algorithm

The empty buffers state $\{\mathbf{n}, \mathbf{m}\} = \{\mathbf{B}, \mathbf{0}\}$ can be reached when the stations happen to be temporarily sluggish. Every Markovian transition probability matrix is, therefore, ergodic with \mathbf{B} as a regeneration point, plus possibly few transient states feeding it as well. This is sufficient to ensure that the above functional equations possess a unique solution with real relative values (Schweitzer 1971, 1984). Schweitzer's iterative solution algorithm is adapted to this problem. The initial step size used in this finite-difference algorithm to ensure fast but stable convergence is

$$\begin{aligned} \tau = \text{Min } [1/\mu_k S + \lambda(\mathbf{B})], & \quad (27) \\ \mathbf{m} \in \phi, & \\ \mathbf{n} \in \Omega, & \\ J \in D(\mathbf{n}, \mathbf{m}). & \end{aligned}$$

A possible initial guess is $v(\mathbf{n}, \mathbf{m}) = 0$ for all $\{\mathbf{n}, \mathbf{m}\} \in \Theta$. The solution algorithm was halted when g could be estimated with $\pm 0.1\%$.

This implementation of the computational algorithm is carried out by a FORTRAN 5 computer program developed especially for this problem. This program generates the optimal policy and computes the desired performance measures. Empirically, the total CPU time is roughly proportional to the size of the state space. For instance, solving a model having 125 states required 12 CPU seconds and 125k bytes of core memory on a CDC 990 computer with the NOS/VE operating system. Convergence was observed after 151 iterations. In general, the computation times increase less than quadratically with the number of states.

We have also developed several measures to accelerate the convergence process. Starting with the step size as given by Equation (27) and once monotone convergence is detected,

we increase it by fourfold and continue doing so until either we get very close to the optimal solution (stable policy and stable g) or the solution diverges. In the latter (very rare) case, we would scale back to the original step size. Other acceleration methods used successfully include the implementation of simple heuristics to generate an initial policy and skipping certain minimization steps at the interior of the state space (no blocking or starvation). This measure was motivated by examining the structural properties of the optimal policy. The frequency of the minimization step is increased adaptively with the convergence of g . On the average, these measures seem to be very effective in reducing the required CPU time by about one order of magnitude for problems with several thousand states.

3. Heuristic policies

In addition to the two optimal policies discussed above, we analyzed four heuristic procedures that attempt at getting close to the optimal system performance. The first three heuristic policies described below are state-dependent, or feedback, control rules and the last one is an open loop, state-independent, rule used as a benchmark procedure.

The following policies were developed and implemented by us:

1. *Fastest Shortest Queue (FSQ)* policy routes parts to the workstation having the minimal number of parts in the input buffer at the decision epoch.

The selection rule used selects type k for processing at the cells such that

$$n_k + m_k \leq n_i + m_i \quad \forall \quad 1 \leq i \leq R$$

and

$$n_k + m_k < B_k .$$

In a case of a tie, the station having the fastest processing rate is selected. Since production rate is maximized (and starvation is minimized) when all stations are busy, it is anticipated that the FSQ will perform well with respect to these objectives.

2. *Work Time Balance (WTB)* policy tries to generate balanced workload contributions at the input buffers of the various workstations. The queue lengths at the stations are scaled by the respective production rates (λ_i) and starvation penalties (c_i), or by the relative value of each part type (w_i) — depending on the objective function in use. This policy attempts at getting equally weighted mean times until starvation at all stations and to reduce the probability of blocking due to excess inventory at any one buffer. It is based on the earlier ideas of Foschini (1977) and Hahne (1981). In our case, however, decisions are made *before* the center starts operating since distinct processing times are required for the various part types at the centers as well as at the stations.

In the case of a minimal starvation objective, the WTB policy selects type k for processing at the centers such that

$$\frac{n_k}{c_k \lambda_k} \leq \frac{n_i}{c_i \lambda_i} \quad \forall \quad 1 \leq i \leq R$$

and

$$n_k + m_k < B_k.$$

The FSQ policy is used in case of a tie.

3. *Weighted Shortest Queue (WSQ)* policy extends the WTB structure to accommodate the current status of both centers and stations designated for each part type. The current inventory value ($= n_i + m_i$) is scaled by the expected transition rate following a selection of a given part type. Higher priority is given to part types associated with lower transition rates.

In the case of a minimal starvation objective, the WSQ policy selects type k for processing at the centers such that

$$\frac{n_k + m_k}{c_k \lambda_k} (\mu(\mathbf{m}, k) + \lambda(\mathbf{n})) \leq \frac{(n_i + m_i)}{c_i \lambda_i} (\mu(\mathbf{m}, i) + \lambda(\mathbf{n})) \quad \forall i, 1 \leq i \leq R$$

and

$$n_k + m_k < B_k.$$

The WTB policy is used in case of a tie.

4. *The Open Loop (OL)* policy is a subset of the WSQ discussed above. It operates in a similar way to the WSQ policy but ignores all the instantaneous system information. Specifically, in a case of minimal starvation objective, the OL policy selects type k for processing at the cells such that

$$\frac{(\mu(\mathbf{m}, k) + \lambda(\mathbf{n}))}{c_k \lambda_k} \leq \frac{(\mu(\mathbf{m}, i) + \lambda(\mathbf{n}))}{c_i \lambda_i} \quad \forall i, 1 \leq i \leq R$$

and

$$n_k + m_k < B_k.$$

The station with the highest value of λ_i is selected in a case of a tie.

The computation of the operational performance measures for these four heuristic procedures is also given in the Appendix.

4. Application examples

The structure of the optimal policy is demonstrated in this section by means of several numerical examples. The first example illustrates the effects of varying the number of centers, while the second example displays the effects of changing the productive capacity of a given set of centers.

4.1. Example set 1

This example illustrates the effects of varying the number of centers on the structure and performance of the optimal part-loading policy. It also demonstrates the wealth of information that can be extracted from the state-space variables. The objective function used here is minimization of the starvation penalties. We consider here an FMS with three stations ($R = 3$) and several centers. The processing rates of these stations are eight, six and four parts per hour ($= \lambda_1, \lambda_2, \lambda_3$). Each station has four buffer spaces ($= B_1 = B_2 = B_3$). These stations are fed by $S (\geq 1)$ parallel centers. Each center processes $21/S$ units per hour ($= \mu_1 = \mu_2 = \mu_3$). The assumed shortage penalties are \$120, \$70, and \$210 per hour idle ($= C_1, C_2, C_3$).

Experimenting with this FMS model, the number of the parallel centers was incremented from one to four ($S = 1, 2, 3$, and 4) in order to examine the interaction between the number of parallel centers and the structure of the optimal policy. The same aggregate productive capacity at the centers, namely 21 units per hour, was retained in these four cases. Tables 1 and 2 present the results of analyzing FMS configurations having two and four centers, respectively. These tables include the system states (n, m) , the state numbers SN, the optimal decisions (M^*, k^*) , and the relative state values $v(n, m)$.

The first line in the Optimal Decision columns in these tables contains the processing plan M^* for the initial (and transient) case. The following lines display, in the same format, the processing plan k^* for all other states. The entries in the State Value columns give the computed values of each state relative to the initial state. From Tables 1 and 2 it is evident that the structure and the size of the state space vary significantly with the number of the FMS centers S . To show this, note for instance that only three states of the centers need to be evaluated in Table 1 when $n = \{2, 2, 1\}$. These states are: SN = 164-166; by contrast, eight states of the centers are evaluated in Table 2 for the same state of the buffers $n = \{2, 2, 1\}$; these eight states are SN = 391-398.

Both tables present sample states drawn from the four mutually exclusive decision regions of the state space as discussed earlier in Section 2.3. These four regions are illustrated below using representative states from Table 2:

1. The initial state

$$\begin{array}{l} |n| = 0, \quad |m| = 0, \\ \text{e.g., } SN = 1 \rightarrow n \{0, 0, 0\}, \quad m \{0, 0, 0\} . \\ \text{Here } |n| = 0, \quad |m| = 0. \end{array}$$

2. Full centers operation

$$0 \leq |n| \leq |B| - S, \quad |m| = S - 1,$$

e.g., SN = 395 \rightarrow $n = \{2, 2, 1\}$, $m = \{1, 1, 1\}$.

Here $|n| = 5$, $|m| = 3$.

3. Partial centers operation with partial blocking

$$|m| + |n| = |B| - 1, \quad 0 \leq m < S - 1,$$

e.g., SN = 540 \rightarrow $n = \{3, 3, 4\}$, $m = \{1, 0, 0\}$.

Here $|n| = 10$, $|m| = 1$.

4. Partial centers operation with complete blocking

$$|m| + |n| = |B|, \quad 0 \leq m \leq S - 1,$$

e.g., SN = 551 \rightarrow $n = \{3, 4, 3\}$, $m = \{1, 0, 1\}$.

Here $|n| = 10$, $|m| = 2$.

It is possible to identify only a few states that appear in both tables (e.g., SN = 252 in Table 1 vs. SN = 540 in Table 2). Such a correspondence in the state structures can only be found in states of partial centers operation with partial or complete blocking (cases 3 or 4 above). It is interesting to note that, in these corresponding states, the same optimal decisions are arrived at for FMS configurations having either two or four centers. The state values, however, are not identical.

In general, the state values $v(n, m)$ represent the *relative* value of one state versus another in the same system. Setting $v(o, o) = 0$ in this FMS model (Equation (22)) means that $v(n, m)$ is the marginal cost of starting the system in state $\{n, m\}$ rather than in state $\{o, o\}$. Following this convention, the blocked cases ($n = B$) should have the smallest relative state values. The positive entries of few states, at the low-buffer end of the state space, indicate that starting the system in any one of these states is even less advantageous than starting it in state $\{o, o\}$.

Several comparative performance measures are displayed in Table 3. The measures presented there are stations' production rate (r_i) stations' utilization (U_i), the utilization of the centers (CU), the centers' effective production rate (CEPR), and the expected starvation cost per time unit (g). These measures refer to four FMS configurations identical to those described above, except for having $S = 1, 2, 3$, and 4 centers. The aggregate productive capacities of the centers in all four systems were identical (i.e., $\mu_i = 21/S$, $i = 1, 2, 3$).

The results presented in Table 3 point at a slight reduction in the FMS performance in spite of the enhanced scheduling flexibility due to the introduction of additional centers. The major reason for this reduction is the blocking control policy restriction that limits the current number of active centers to the total number of available buffer spaces. Such a restriction means that an increase in S will cause a decrease in the productivity of the centers. This inverse relationship is particularly pronounced at the high-buffer end of the state space. These data lend further support to a well-known conclusion in queuing theory that an isolated server is preferable to several parallel servers having the same aggregate processing capabilities.

Table 3. Summary of performance measures for a varying number of parallel centers ($\mu_i = 21/S$, $i = 1, 2, 3$), and three stations

Performance Measure	i	Number of parallel centers			
		$S = 1$	$S = 2$	$S = 3$	$S = 4$
r_i	1	7.13	7.03	6.95	6.86
	2	5.90	5.87	5.81	5.73
	3	3.95	3.93	3.92	3.90
U_i	1	89.22%	87.93%	86.89%	85.85%
	2	98.44%	97.99%	96.92%	95.53%
	3	98.81%	98.26%	98.01%	97.54%
CU		80.94%	80.21%	79.46%	78.57%
CEPR		16.98	16.83	16.68	16.49
g		21.18	25.57	31.26	238.69

4.2. Example set 2

This example demonstrates several attributes of the interaction between the processing capacity of the centers in relation to that of the stations and the structure of the optimal policy. The interaction with the FMS performance is also demonstrated. The FMS configuration considered here is similar to the one analyzed in Example 1: $R = 3$, $S = 3$, $\mathbf{B} = \{4, 4, 4\}$, $\lambda_1 = 8$, $\lambda_2 = 6$, $\lambda_3 = 4$, $C_1 = 120$, $C_2 = 370$, and $C_3 = 210$. Five levels of processing capacities at the centers are evaluated here. Specifically, the values assumed by μ_i ($i = 1, 2, 3$) are 3, 5, 7, 9, and 11 parts/hour. Table 4 gives the results of computing the optimal processing decisions and the relative state values for two extreme cases. The first assumed $\mu_1 = \mu_2 = \mu_3 = 3$ (*underpowered* centers), and the second case assumed $\mu_1 = \mu_2 = \mu_3 = 11$ (*overpowered* centers).

The results presented in Table 4 prove that the optimal policy provides for preferential scheduling of certain part types with the objective of minimizing the weighted starvation penalty incurred by the stations when they run out of input parts. This preferential scheduling pattern is clearly pronounced in cases where all admissible *stations* are the *same* state. For example, consider $SN = 330$ in Table 4: stations 1 and 3 are in the same state (i.e., $n_1 = n_3$, $m_1 = m_3$) and station 2 is blocked ($n_2 = B_2$). The optimal decision at this state (= 001) demonstrates the priority of part type 3 over 1. The priority of part type 2 over 3 is demonstrated in state $SN = 459$ for the two extreme cases of the centers processing rate capacity. Several performance measures computed for three values of μ_i are presented in Table 5.

The relative changes in the magnitude of U_i and CU, as a function of μ_i , are also depicted by Figure 3. This figure demonstrates the significance of capacity balancing and planning in these buffered systems: when $\mu_i = 3$, the centers are the bottleneck (i.e., CU approaches 100%); as the centers capacity increases, the stations quickly become the bottleneck resource.

Table 4. Optimal processing decisions and the relative state values for an FMS having six work cells: three centers and three stations

State number SN	State space		$\mu_i = 3$ ($i = 1, 2, 3$) (underpowered)	$\mu_i = 11$ ($i = 1, 2, 3$) (overpowered)
	n	m	Optimal decision $M^*, k(^*)$	State value $v(n, m)$
1	0 0 0	0 0 0	030	0.00
2	0 0 0	0 0 2	010	4.39
3	0 0 0	0 1 1	010	1.04
4	0 0 0	0 2 0	010	0.00
5	0 0 0	1 0 1	010	32.33
67	0 2 2	0 1 1	010	-168.93
68	0 2 2	0 2 0	001	-168.93
69	0 2 2	1 0 1	010	-153.72
70	0 2 2	1 1 0	010	-155.44
71	0 2 2	2 0 0	010	-134.79
72	0 2 3	0 1 1	010	-192.46
73	0 2 3	0 2 0	001	-192.46
74	0 2 3	1 0 1	010	-182.92
75	0 2 3	1 1 0	010	-185.85
76	0 2 3	2 0 0	010	-170.58
77	0 2 4	0 2 0	100	-208.66
78	0 2 4	1 1 0	010	-208.66
79	0 2 4	2 0 0	010	-199.16
80	0 3 0	0 0 2	001	-133.10
218	1 4 2	0 0 2	100	-228.43
219	1 4 2	1 0 1	001	-228.43
220	1 4 2	2 0 0	001	-220.95
221	1 4 3	1 0 1	100	-245.44
330	2 4 2	1 0 1	001	-243.84
331	2 4 2	2 0 0	001	-236.38
458	4 2 2	0 0 2	010	-227.20
459	4 2 2	0 1 1	010	-229.69
460	4 2 2	0 2 0	001	-229.69
480	4 4 3	0 0 0	001	-292.45
481	4 4 3	0 0 1	000	-292.45
482	4 4 4	0 0 0	000	-305.57

Table 5. Summary of performance measures for an FMS having six work cells (three centers and three stations) and varying processing rates at the centers

Performance measure	i	Processing capacity of the centers		
		$\mu_i = 3$ ($i = 1, 2, 3$)	$\mu_i = 7$ ($i = 1, 2, 3$)	$\mu_i = 11$ ($i = 1, 2, 3$)
r_i	1	1.41	6.95	7.87
	2	4.59	5.81	5.96
	3	2.95	3.92	3.99
U_i	1	17.63%	86.89%	98.46%
	2	76.62%	96.94%	99.47%
	3	73.88%	98.01%	99.83%
CU		99.61%	79.41%	54.05%
CEPR		8.95	16.68	17.82
g		240.17	31.26	4.16

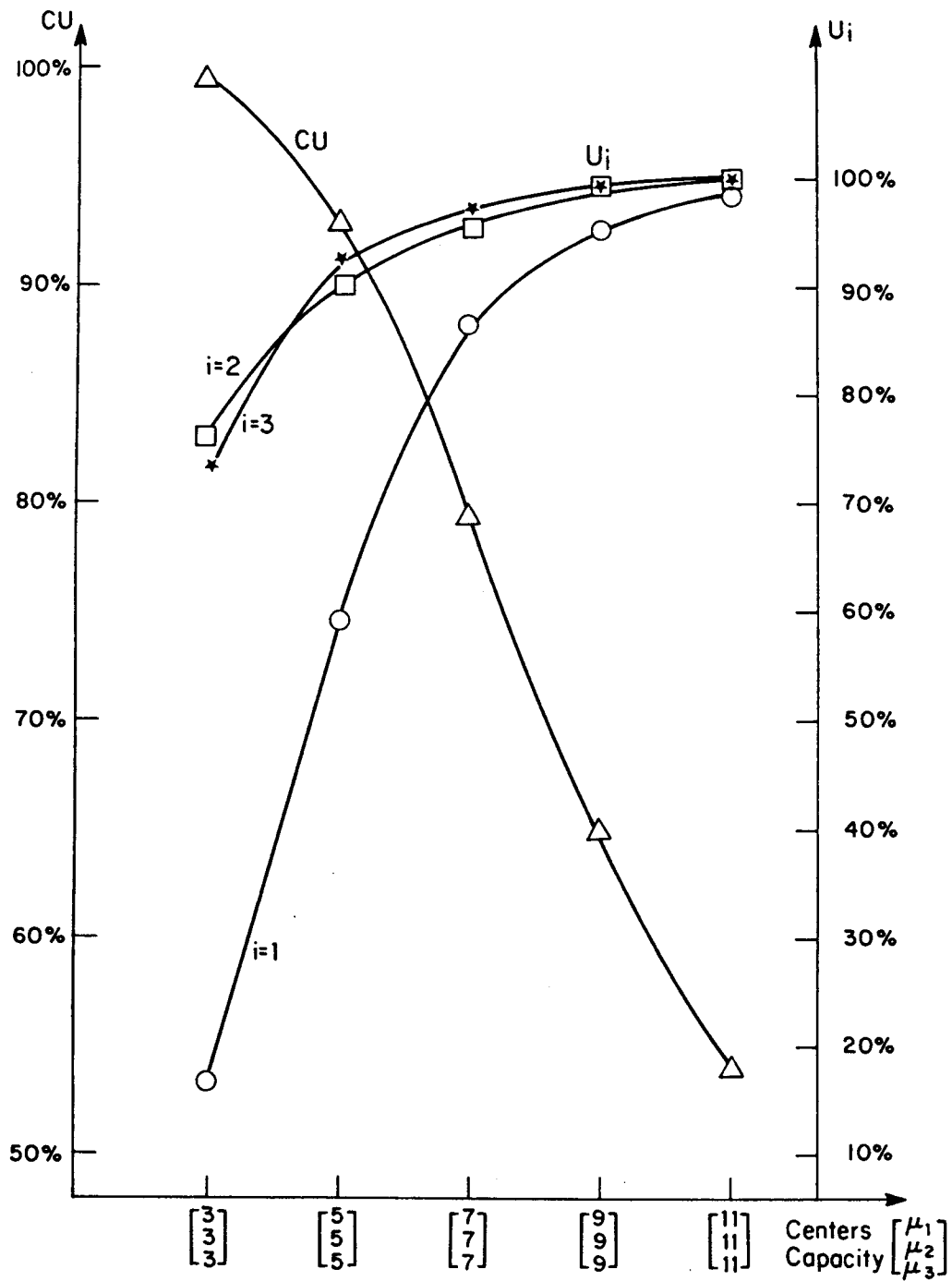


Figure 3. The mean center utilization (CU) and the station utilization (U_i) as a function of the centers' capacity (μ).

4.3. Example set 3

This example compares the optimal throughput maximizing policy with the four heuristic policies described above. The system considered has six work cells: two centers ($S = 2$) and four stations ($R = 4$). The processing rates of the centers are 16, 4, 2, and 8 parts per hour ($= \mu$) for part types one to four, respectively. The processing rates and the buffer allocations for this example set are depicted in Table 6. Table 7 presents a summary of the weighted throughput rates for the optimal and the heuristic policies. The best overall heuristic performance is scored by the WSQ policies, with the WTB and OL in a close tie for the worst. These results indicate that the WSQ policy, which uses state-dependent FMS information regarding the status of both centers and stations, comes fairly close to the optimal performance. Such performance is for most practical purposes almost as good as the optimal policy and has the great advantage of a negligible computational effort.

A typical summary of the FMS performance measures for sample case V is given by Table 8. It points to the fact that the WSQ policy tends to generate a load profile that is somewhat similar to that of the optimal policy. In general, the OL and WTB policies were found to generate less items (e.g., CEPR) and less revenues as they overassign production of parts for lines two and four.

Table 6. Buffer allocation and the stations processing rates (in units/hour) data as used in the six sample cases of example set 3

Sample case	Buffer allocation : B_i				Stations processing rates : λ_i			
	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 1$	$i = 2$	$i = 3$	$i = 4$
I	2	3	3	2	16	4	2	8
II	2	3	3	2	8	2	1	4
III	2	3	3	2	12	3	1.5	6
IV	3	4	4	3	8	2	1	4
V	3	4	4	3	16	4	2	8
VI	3	4	4	3	12	3	1.5	6

Table 7. The weighted throughput rate ($= g$) for the optimal and the four heuristic policies

Sample case	Policy				
	Optimal	OL	FSQ	WTB	WSQ
I	265.53	253.88	253.88	246.10	263.38
II	202.76	200.75	200.75	199.11	202.20
III	274.78	236.14	236.14	231.83	240.5
IV	244.67	224.51	239.17	226.06	243.86
V	326.02	229.33	271.50	218.13	305.01
VI	290.01	231.72	264.44	230.10	282.56
(average) \bar{g}	267.26	229.38	244.31	225.22	256.26
$\frac{\bar{g}}{\bar{g}^*} 100$	—	85.82%	91.40%	84.26%	95.86%

Note: The last row in the table presents the relative percentage deviation from the average optimal value, \bar{g}^* .

Table 8. Summary of the FMS performance measures for sample case *V*

Performance measure	<i>i</i>	Policy				
		Optimal	OL	FSQ	WTB	WSQ
r_i	1	7.68	2.99	5.21	2.52	6.94
	2	1.85	1.95	1.74	1.85	0.74
	3	1.89	1.36	1.49	1.26	1.47
	4	1.70	2.31	2.14	2.43	2.47
u_i	1	48.01%	18.69%	32.57%	15.81%	43.43%
	2	46.47%	48.79%	43.63%	46.37%	18.51%
	3	94.88%	68.11%	74.92%	63.05%	73.53%
	4	21.26%	28.91%	26.79%	30.47%	30.88%
CU		99.56%	99.57%	99.84%	99.81%	99.89%
CEPR		13.13	8.62	10.59	8.08	11.63
g		326.02	229.33	271.50	218.13	305.01

5. Summary and conclusions

This article presents the mathematical formulations for generating two optimal and four heuristic policies for dynamic load controls in FMSs with stochastic processing rates. The objective functions considered deal with the minimization of the weighted throughput rates. In general, the real-time scheduling of FMSs is a complicated process in which the control logic should consider a broad spectrum of instantaneous state variables and the expected future impact of each decision at each time epoch. In an effort to better understand these real-time part-loading issues, we present an analytic state-transition framework that considers major interaction factors between the work centers of a flexible glass lens manufacturing facility. This framework is extended into several optimal and heuristic control policies, and its applicability is illustrated by several numerical examples.

A close examination of various examples, computed for an extensive range of problems, seems to indicate that the optimal policy attempts to evade the risk of starvation at those stations that have the highest expected revenue generation rates. The optimal policy, therefore, generates hedging (or safety) buffer stocks of parts in these stations before it starts to process parts for stations associated with lower expected revenue generation rates. This is clearly depicted by the underpowered configuration we studied. In this case it was preferred to reduce the risk of starvation at certain stations as compared to reducing the concurrent starvation penalty of others.

The required level of these hedging stocks is a function of the processing capacity (or the response time) of the centers relative to that of the stations. Increasing the relative processing capacity of the centers (or reducing their response time) is expected to lead to a reduction in the required levels of these hedging stocks. An increase in the relative processing rates of additional part types tends to have only secondary priority in terms of their expected starvation penalty. Indeed, such a change in the pattern of the optimal

decision policy was also demonstrated. The relative state values in the overpowered case are significantly *larger* than their counterparts in the underpowered case. This means that an FMS with overpowered centers reduces the relative importance of starting its operation with full buffers.

Several heuristic policies are also developed and examined. We show that certain heuristic procedures come fairly close to the optimal performance. These feedback control policies are almost as good as the optimal policies and have the clear advantage of requiring a minimal computational effort. While a large number of studies support the robustness of the assumptions used in our study, we see a need for further research into such issues as processing times characterization, sensitivity to the operational modes of the transporters, and evaluating detailed application experiences in other industrial systems. Current research efforts deal with further evaluation of the management policies developed here under stationary and transient workloads.

Appendix A: Operational performance measures

A.1. Definitions

Define by \mathbf{M}^* the decision vector for state $(\mathbf{n} = \mathbf{o}, \mathbf{m} = \mathbf{o})$ and by $k^* \triangleq k(\mathbf{n}, \mathbf{m})$ the optimal decisions for all other states. With an optimal or heuristic decision vector, it becomes feasible to compute the FMS performance measures as a function of \mathbf{M}^* and k^* . The detailed equations for evaluating several major performance measures are developed here. These equations can be used for *both* objective functions and for any heuristic rule as well. For a general scheme of evaluating additional measures, see Howard (1971), Dreyfus and Law (1977), or Seidmann and Tenenbaum (1989).

A.2. The throughputs of the stations

Let r_i ($i = 1, 2, \dots, R$) denote the production rate of station i following the optimal policy. Next, define a Semi-Markovian process producing a fractional reward $Q(\mathbf{n}, \mathbf{m})$ equal to the probability that a part type i will be completed at the centers by the next decision epoch:

$$Q(\mathbf{n}, \mathbf{m}) = \begin{cases} \mu_i M_i / \mu(\mathbf{M}^*) & \text{if } S \geq 1, |\mathbf{n}| = 0, |\mathbf{m}| = 0, \\ \mu(m_i, k^*) / \mu(\mathbf{m}, k^*) & \text{if } S \geq 1, |\mathbf{m}| = S - 1, \\ & 0 \leq |\mathbf{n}| \leq |\mathbf{B}| - S, \\ & \{\mathbf{n}, \mathbf{m}\} \in \Theta, \\ \mu(m_i, k^*) / (\mu(\mathbf{m}, k^*) + \lambda(\mathbf{n})) & \text{if } S > 1, 0 \leq |\mathbf{m}| < S - 1, \\ & |\mathbf{m}| + |\mathbf{n}| = |\mathbf{B}| - 1, \\ & \{\mathbf{n}, \mathbf{m}\} \in \Theta, \\ m_i \mu_i / (\mu(\mathbf{m}) + \lambda(\mathbf{n})) & \text{if } S \geq 1, |\mathbf{m}| + |\mathbf{n}| = |\mathbf{B}|, \\ & 0 \leq |\mathbf{m}| \leq S - 1, \\ & \{\mathbf{n}, \mathbf{m}\} \in \Theta. \end{cases} \quad (\text{A.1})$$

The value equations for computing the production rate of station i are

$$v(\mathbf{o}, \mathbf{o}) = Q(\mathbf{n}, \mathbf{m}) - r_i/\mu(\mathbf{M}^*) + \sum_{i=1}^R [M_i \mu_i/\mu(\mathbf{M})] v(\mathbf{e}^i, \mathbf{M} - \mathbf{e}^i) \quad (\text{A.2})$$

$$\text{if } S > 1, \quad |\mathbf{n}| = 0, \quad |\mathbf{m}| = 0,$$

$$v(\mathbf{n}, \mathbf{m}) = Q(\mathbf{n}, \mathbf{m}) - r_i/\mu(\mathbf{m}, k^*)$$

$$+ \sum_{i=1}^R [\mu(m_i, k^*)/\mu(\mathbf{m}, k^*)] \sum_{\mathbf{o} \leq \mathbf{j} \leq \mathbf{n}} [P(\mathbf{j} | \mathbf{n}, \mu(\mathbf{m}, k^*))] v(\mathbf{n} - \mathbf{j} + \mathbf{e}^i, \mathbf{m} + \mathbf{e}^{k^*} - \mathbf{e}^i) \quad (\text{A.3})$$

$$\text{if } \{\mathbf{n}, \mathbf{m}\} \in \Theta, \quad S \geq 1, \quad |\mathbf{m}| = S - 1,$$

$$0 \leq |\mathbf{n}| \leq |\mathbf{B}| - S,$$

$$v(\mathbf{n}, \mathbf{m}) = Q(\mathbf{n}, \mathbf{m}) - r_i/(\mu(\mathbf{m}, k^*) + \lambda(\mathbf{n})) \quad (\text{A.4})$$

$$+ \sum_{i=1}^R [\mu(m_i, k^*)/(\mu(\mathbf{m}, k^*) + \lambda(\mathbf{n}))] v(\mathbf{n} + \mathbf{e}^i, \mathbf{m} + \mathbf{e}^{k^*} - \mathbf{e}^i)$$

$$+ \sum_{i \in A(\mathbf{n})} [\lambda_i/(\mu(\mathbf{m}, k^*) + \lambda(\mathbf{n}))] v(\mathbf{n} - \mathbf{e}^i, \mathbf{m} + \mathbf{e}^{k^*})$$

$$\text{if } \{\mathbf{n}, \mathbf{m}\} \in \Theta, \quad S > 1, \quad 0 \leq |\mathbf{m}| < S - 1,$$

$$|\mathbf{m}| + |\mathbf{n}| = |\mathbf{B}| - 1,$$

$$v(\mathbf{n}, \mathbf{m}) = Q(\mathbf{n}, \mathbf{m}) - r_i/(\mu(\mathbf{m}) + \lambda(\mathbf{n})) \quad (\text{A.5})$$

$$+ \sum_{i=1}^R [\mu_i m_i/(\mu(\mathbf{m}) + \lambda(\mathbf{n}))] v(\mathbf{n} + \mathbf{e}^i, \mathbf{m} - \mathbf{e}^i)$$

$$+ \sum_{i \in A(\mathbf{n})} [\lambda_i/(\mu(\mathbf{m}) + \lambda(\mathbf{n}))] v(\mathbf{n} - \mathbf{e}^i, \mathbf{m})$$

$$\text{if } \{\mathbf{n}, \mathbf{m}\} \in \Theta, \quad S \geq 1, \quad 0 \leq |\mathbf{m}| \leq S - 1,$$

$$|\mathbf{m}| + |\mathbf{n}| = |\mathbf{B}|,$$

$$v(\mathbf{o}, \mathbf{o}) = 0. \quad (\text{A.6})$$

These values equations are solved by the iterative algorithm described in Section 2.6. The long-run expected reward per unit time (the *gain rate*) r_t , is the effective throughput rate of station t .

A.3. The utilization of the stations

Given r_t , the utilization of station t is given by

$$U_t = r_t/\lambda_t. \quad (\text{A.7})$$

A.4. The effective production rate of the centers

The effective production rate of the centers is computed by

$$\text{CEPR} = \sum_{t=1}^R r_t. \quad (\text{A.8})$$

A.5. The utilization of the centers

The transitions in the system states start and terminate at the end-of-processing event either at the centers or at one of the stations. The utilization of the S parallel centers, therefore, cannot be computed directly from the stations throughputs. To obtain this measure, we let CU denote the expected utilization of the centers. To compute CU, define another Semi-Markovian process similar to the one used for getting the throughput rate of the stations. In this case, however, the following one transition reward is used:

$$Q(\mathbf{n}, \mathbf{m}) = \begin{cases} | \mathbf{M} | / (\mu(\mathbf{M}^*) S) & \text{if } S > 1, | \mathbf{n} | = 0, | \mathbf{m} | = 0, \\ (| \mathbf{m} | + 1) / (\mu(\mathbf{m}, k^*) S) & \text{if } S \geq 1, | \mathbf{m} | = S - 1, \\ & 0 \leq | \mathbf{n} | \leq | \mathbf{B} | - S, \\ & \{ \mathbf{n}, \mathbf{m} \} \in \Theta, \\ \\ (| \mathbf{m} | + 1) ((\lambda(\mathbf{n}) + \mu(\mathbf{m}, k^*) S)) & \text{if } S > 1, 0 \leq | \mathbf{m} | < S - 1, \\ & | \mathbf{m} | + | \mathbf{n} | = | \mathbf{B} | - 1, \\ & \{ \mathbf{n}, \mathbf{m} \} \in \Theta, \\ \\ | \mathbf{m} | / ((\lambda(\mathbf{n}) + \mu(\mathbf{m})) S) & \text{if } S \geq 1, | \mathbf{m} | + | \mathbf{n} | = | \mathbf{B} |, \\ & 0 \leq | \mathbf{m} | \leq S - 1, \\ & \{ \mathbf{n}, \mathbf{m} \} \in \Theta. \end{cases} \quad (\text{A.9})$$

This immediate reward $Q(\mathbf{n}, \mathbf{m})$ is simply equal to the average number of activity hours accumulated, per center, by all the centers between two consecutive decision epochs $t = \text{ratio of active centers times the expected duration of the corresponding time interval}$.

The gain rate of this Semi-Markovian process is CU (rather than r_i in Equations (A.2)–(A.6)); its value is revealed by solving these functional value equations. Clearly, (1-CU) is that fraction of the centers' processing power that is lost due to blocking at the buffers.

Acknowledgments

The authors wish to thank the referees, the associate editor, and the editor for many valuable comments. The authors would also like to thank Mrs. Deborah Neulander, from the School of Engineering at Tel Aviv University, for her help in developing the scientific computer codes and in conducting the numerical experiments.

References

- Akella, R., Choong, Y., and Gershwin, S.B., "Real-Time Production Scheduling of an Automated Cardline," *Annals of Operations Research*, Vol. 3, pp. 403–425 (1985).
- Alam, M., Gupta, D., Ahmad, S.I., and Raouf, A., "The Performance Modeling and Evaluation of Flexible Manufacturing Systems Using Semi-Markov Approach," in *Flexible Manufacturing: Recent Developments in FMS, Robotics, CAD/CAM, CIM*, A. Raouf and S.I. Ahmad (Eds.), Elsevier Science Publishers B.V., Amsterdam, pp. 87–118 (1985).
- Buzacott, J.A. and Hanifin, L.E. "Models of Transfer Lines with Inventory Banks—A Review and Comparison," *AIIE Transactions*, Vol. 10, No. 2, pp. 197–209 (1978).
- Cinlar, E., "Decomposition of a Semi-Markovian Process Under a State Dependent Rule," *Journal of Applied Mathematics*, Vol. 15, No. 2, pp. 252–263 (1967).
- Costa, A. and Garetti, M., "Design of a Control System for a Flexible Manufacturing Cell," *Journal of Manufacturing Systems*, Vol. 4, No. 1, pp. 65–84 (1985).
- Drozda, T.J. (Ed.), *Flexible Manufacturing Systems, SME*, Dearborn, MI (1988).
- Dreyfus, S.E. and Law, A.M., *The Art and Theory of Dynamic Programming*, Academic Press, New York (1988).
- Elsayed, E.A. and Hwang, C.C. "Analysis of Two-Stage Manufacturing Systems with Buffer Storage and Redundant Machines," *International Journal of Production Research*, Vol. 24, No. 1, pp. 187–201 (1986).
- Foschini, G.J., "On Heavy Traffic Diffusion Analysis and Dynamic Routing in Packet Switched Networks," *Computer Performance*, K.M. Chandy and M. Reiser (Eds.), North Holland Publishing Company, New York, pp. 499–513 (1977).
- Gershwin, S.B. and Berman, O., "Analysis of Transfer Lines Consisting of Two Unreliable Machines With Random Processing Times and Finite Storage Buffers," *AIIE Transactions*, Vol. 13, No. 1, pp. 2–11 (1981).
- Gershwin, S.B., "Representation and Analysis of Transfer Lines With Machines That Have Different Processing Rates," MIT Laboratory for Information and Decision Systems, Report No. LIDS: 35–433 (1985).
- Gershwin, S.B., Hildebrant, R.R., Suri, R. and Mitter, K., "Control Perspective on Recent Trends in Manufacturing Systems," *IEEE Control Systems Magazine*, Vol. 6, No. 2, pp. 3–15 (1986).
- Gray, A., Seidmann, A. and Stecke, K.E., "Tool Management in Automated Manufacturing: Operational Issues and Decision Problems," Working Paper No. CMOM 88–03, William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, New York (1988).
- Hahne, E.L., "Dynamic Routing In An Unreliable Manufacturing Network With Limited Storage," MIT Laboratory for Information and Decision Systems, Report No. LIDS-TH-1063 (1981).
- Han, M.H. and McGinnis, L.F., "Throughput Rate Maximization in Flexible Manufacturing Cells," *IIE Transactions*, Vol. 20, No. 4, pp. 409–412 (1988).
- Hildebrant, R.R., "Scheduling Flexible Machining Systems Using Mean Value Analysis," Proceedings of the IEEE Conference on Decision and Control, Albuquerque, NM, pp. 701–706 (1980).
- Howard, R.A., *Dynamic Probabilistic Systems, Vol. II: Semi-Markov and Decision Processes*, John Wiley, New York (1971).

- Jewell, W.J., "Markov-Renewal Programming: I and II," *Operations Research*, Vol. 11, No. 6, pp. 938-948, 949-971 (1963).
- Kusiak, A., *Modelling and Design of Flexible Manufacturing Systems*, Elsevier, Amsterdam (1986).
- Lin, W. and Kumar, P.R. "Optimal Control of Queuing System With Two Heterogeneous Servers," *IEEE Transactions Automatic Control*, Vol. 27, No. 2, pp. 696-703 (1984).
- Masri, S. and Hausman, V.H., "Dynamic Scheduling in a Flexible Manufacturing System with Failure Prone Machines," Working Paper, Department of Industrial Engineering and Engineering Management, Stanford University, Stanford, CA (1988).
- Nof, S.Y. (Ed.) *The Handbook of Industrial Robotics*, Wiley, New York (1985).
- Okamura, K. and Yamashina, H., "Justification For Installing Buffer Stocks in Unbalanced Two Stage Automatic Transfer Lines," *AIIE Transactions*, Vol. 11, No. 4, pp. 308-312 (1979).
- Pinedo, M.L., Wolf, B. and McCormick, S.T., "Sequencing in a Flexible Assembly Line with Blocking to Minimize Cycle Time," *Proceedings of the Second ORSA/TIMS Conference on Flexible Manufacturing Systems: Operations Research Models and Applications*, Elsevier Science Publishers B.V., Amsterdam, pp. 499-508 (1986).
- Rachamadugu, R. and Stecke, K.E., "Classification and Review of FMS Scheduling Procedures," Working Paper, University of Michigan, Ann Arbor, MI (1989).
- Schweitzer, P.J., "Iterative Solution Of The Functional Equations Of Undiscounted Markov Renewal Programming," *Journal of Mathematical Analysis and Applications*, Vol. 34, pp. 495-501 (1971).
- Schweitzer, P.J., "On The Existence Of Relative Values For Undiscounted Markovian Decision Processes With a Scalar Gain Rate," *Journal of Mathematical Analysis and Applications*, Vol. 104, No. 1, pp. 67-78 (1984).
- Seidmann, A. and Schweitzer, P.J., "Part Selection Policy For A Flexible Manufacturing Cell Feeding Several Production Lines," *IIE Transactions*, Vol. 16, No. 4, pp. 355-362 (1984).
- Seidmann, A. and Tenenbaum, A., "A Computational Approach to Optimal Queuing Systems Controls with Finite Buffers and With Multiple Component Cost Functions," *IEEE Transactions on Systems, Man, and Cybernetics* (1989).
- Shalev-Oren, S., Seidmann, A., and Schweitzer, P.J., "Analysis of Flexible Manufacturing Systems with Priority Scheduling: PMVA," *Annals of Operations Research*, Vol. 3, pp. 115-139 (1985).
- Smith, M.L., Ramesh, R., Dudek, R.A., and Blair, E.L., "Characteristics of U.S. Flexible Manufacturing Systems—A Survey," *Proceedings of the Second ORSA/TIMS Conference on Flexible Manufacturing Systems: Operations Research Models and Applications*, Elsevier Science Publishers B.V., Amsterdam, pp. 477-486 (1986).
- Stecke, K.E., and Suri, R., (Eds.), *Proceedings Of The Second ORSA/TIMS Special Interest Conference on Flexible Manufacturing Systems*, Elsevier Science Publishers, B.V., Amsterdam, pp. 477-486 (1986).
- Stecke, K.E., "Algorithms to Efficiently Plan and Operate a Particular FMS," *International Journal of Flexible Manufacturing Systems*, Vol. 1, No. 4, pp. 287-324 (1989).
- Stidham, S. and Altiock, R., "Production Lines With Unreliable Machines and Finite Buffers," paper presented at the ORSA/TIMS Conference, San Francisco, CA (1984).
- Suri, R. and Diehl, G.W., "A Variable Buffer Size Model and Its Use In Analyzing Closed Queuing Networks With Blocking," *Management Science*, Vol. 32, No. 2, pp. 206-224 (1986).
- Suri, R. and Whitney, C.R., "Decision Support Requirements In Flexible Manufacturing," *Journal of Manufacturing Systems*, Vol. 3, No. 1, pp. 61-69 (1984).
- Yao, D.D. and Buzacott, J.A., "Models of Flexible Manufacturing Systems With Local Buffers," *International Journal of Production Research*, Vol. 24, No. 1, pp. 107-117 (1986).
- Yao, D.D. and Buzacott, J.A., "Modeling a Class Of State Dependent Routing In Flexible Manufacturing Systems," *Annals of Operations Research*, Vol. 3, pp. 153-167 (1985).
- Yao, D.D. and Shanthikumar, J.G., "The Optimal Input Rate To A System Of Manufacturing Cells," *INFOR*, Vol. 25, pp. 57-65 (1987).

Abraham Tenenbaum obtained his B.S. degree in Industrial and Management Engineering in 1983 from Ben-Gurion University and a Masters in Industrial Engineering (Cum Laude) from Tel-Aviv University in 1986. He has been involved in developing large scale stochastic models for performance evaluation of manufacturing, computers, and logistical systems. Currently he is with the Management Information Systems group at the world head office of Bank Hapoalim, Ltd. in Tel-Aviv, Israel.

N. Viswanadham is currently a Professor in the Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India. His current research interests are in the areas of automated manufacturing systems and fault-tolerant control system design, and he has authored several papers in these areas. He is the co-author of the book *Reliability of Computer and Control System* published by North-Holland. He is a member of the editorial boards of several international journals.