

OPTIMAL DYNAMIC ROUTING IN FLEXIBLE MANUFACTURING SYSTEMS WITH LIMITED BUFFERS

Abraham SEIDMANN

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

Abstract

An optimal routing policy is obtained for Flexible Manufacturing Systems (FMSs) with limited buffers at the work stations. This policy is used to effectively drive a robotic material handling system. The routing decisions are made by a supervising computer on a real-time basis in order to avoid any work station running out of inputs and to control the blocking of the material handling system. Using our model, general material handling times can be assumed. The optimal policy and several key performance measures are computed, following the problem formulation as a continuous-time, semi-Markovian decision process. Fast convergence and computational stability are ensured by the ergodic solution algorithm augmented to solve the functional equations of the renewal process. The solution algorithm was implemented, tested on an extensive range of problems regarding the structure and the performance of the optimal policy. Complex environments involving diverse processing times, as well as very limited buffer storage, were examined. The interaction between the allocation of buffer spaces to work stations, the structural properties of the optimal monotone (threshold-type) policy and the system performance are also investigated.

Keywords

Flexible manufacturing system, dynamic routing, material handling, robotics, semi-Markovian decision process, optimal control, stochastic optimization.

1. Introduction

In a flexible manufacturing system (FMS), individual part movements are made feasible by the automated transportation system. All the operations carried out, both at the work stations (including parts and tool changes) and by the material handling system, are entirely under real-time computer control. The environment of an FMS is, therefore, completely different from that of a conventional job shop [1]: it provides additional production capabilities on the one hand, while on the other hand it imposes

special constraints on the scheduling function, which should be adapted accordingly. Previous studies of this problem have been concerned mainly with the prediction of the FMS throughput, under various flow control policies. Later, several network-of-queues models have been developed by Solberg [29], Hildebrant [13], Suri [32], Stecke and Solberg [31], Shalev-Oren et al. [27], and others [4,17,20]. Such models were used to explore the impact of various workload allocation patterns and of grouping machine assignments to the FMS work stations (see, e.g. [31]).

Markovian decision models were used by Hahne [11] and by Seidmann and Schweitzer [24] to develop part selection policies. Other relevant studies and design issues were recently presented by Buzacott and Yao [3], Dupont-Gatelmard [6], Stecke [30], Yao and Shanthikumar [38], and by Gershwin et al. [10]. Most of these studies have assumed unlimited local buffers at the work stations. Many industrial FMS installations, however, currently operate with very small local buffers at the machines [2,6,8,12,15].

Of the few studies analyzing FMSs with state-dependent routing and limited local buffers, those of Yao and Buzacott [35,36] seem to be the most directly related to the present study. They considered (among other things) the probabilistic shortest queue (PSQ) scheme for an FMS with limited local buffers at each station. Their formal FMS model is similar to the one detailed in sect. 2 of this paper. These models describe an FMS operating with a central material handling device (i.e. rack and pinion carts) which distributes the pallets and parts from a central storage buffer to the work stations. A "return conveyor" carries the processed parts with their pallets back to the central buffer. Hatvany [12] presents several sample systems operating with these devices at Toyoda Koki (Okasaki Works), at Toyota Motors, and at Fujitsu Fanuc (Hino Works).

The optimal routing policy for a generic FMS model similar to the one discussed above is developed in this paper. Routing policies are computed using a semi-Markovian decision process. While the probabilistic shortest queue scheme was used in [35,36] as an analytical surrogate for the shortest queue heuristic, this paper presents an efficient scheme for recursively generating the *optimal* routing policy. We further extend these earlier studies by formulating an improved optimal production control strategy which permits *temporary suspension* of the material handling system at certain instances, even though the input buffers at the stations are not full. Employing this control strategy leads to improved FMS performance despite possible reduction in the utilization of the material handling system. Our extended formulation also allows for *general material handling times*. Several closed form expressions are developed for the case where the material handling times have an Erlang distribution; hence, the entire spectrum of process variability parameters – from exponential to deterministic – can be explored. Finally, new analytic formulations for the explicit computation of several key *performance measures* such as the number of starvation periods, the expected duration of each starvation period, or the mean buffer occupancies at the decision epochs are presented.

The paper is organized as follows: the FMS model is formulated and its functional equations are described in sect. 2. Several performance measures for evaluating the performance of the system under a given routing scheme are formulated and computed in sect. 3. Numerical examples are presented in sect. 4. Section 5 concludes the paper.

2. The model

2.1. FACILITY STRUCTURE

The FMS to be studied is conceptualized as having a fixed number of parts N , circulating throughout the system in accordance with specified routing requirements. This number represents the number of pallets or fixtures available in the plant. A centralized material handling station (MHS) links together all the M (≥ 1) work stations (e.g. milling and turning centers, gear shapers, etc.). Parts refixturing, loading and unloading (onto and from the pallets), and all other preparatory activities are conducted at the MHS [18].

Work station (machine) k , $1 \leq k \leq M$, and its input buffers have room for $B_k \geq 1$ parts, one in processing plus up to $B_k - 1$ in the buffers. Work stations may be starved (idle), but they are never jammed, since they are the most expensive elements in the FMS. To avoid jamming the work stations, a return conveyor immediately clears away the finished parts from the work stations and carries them back to the central buffer of the MHS [12,15]. It is assumed that there are at least B_k parts available for work station k in order to ensure proper utilization of the FMS [29,36]. Following [35], the time delay of parts on the return conveyor is ignored in this model. This is a reasonable assumption, since in all states, except for the blocked case where no decision is made, the centralized material handling station has at least one part (see sect. 2.4). Figure 1 depicts the physical layout of two work stations where the pallet magazines handle the local input buffers ([12], pp. 51–56).

The time z required for the MHS to route a part to its next-station k , as well as for the necessary preparatory work, has an *arbitrary distribution* $G_k(z)$. This is tractable because the decision epochs occur, in the renewal setting, only when a delivery is completed or when a station completes a part processing and the MHS is idle. Specialized analytical expressions are given for the case in which G_k is an *Erlang distribution* with L_k stages ($L_k = 1, 2, \dots, \infty$) and a mean of $1/\mu_k$ (see eq. (1)):

$$dG_k(z) = \frac{(L_k \mu_k)^{L_k} z^{L_k-1}}{(L_k - 1)!} e^{-L_k \mu_k z} dz, \quad z \geq 0, \quad L_k = 1, 2, \dots \quad (1)$$

Notice that if $L_k = 1$, the above p.d.f. is reduced to the exponential p.d.f. This accounts for the time required by the MHS to traverse from its current location to the

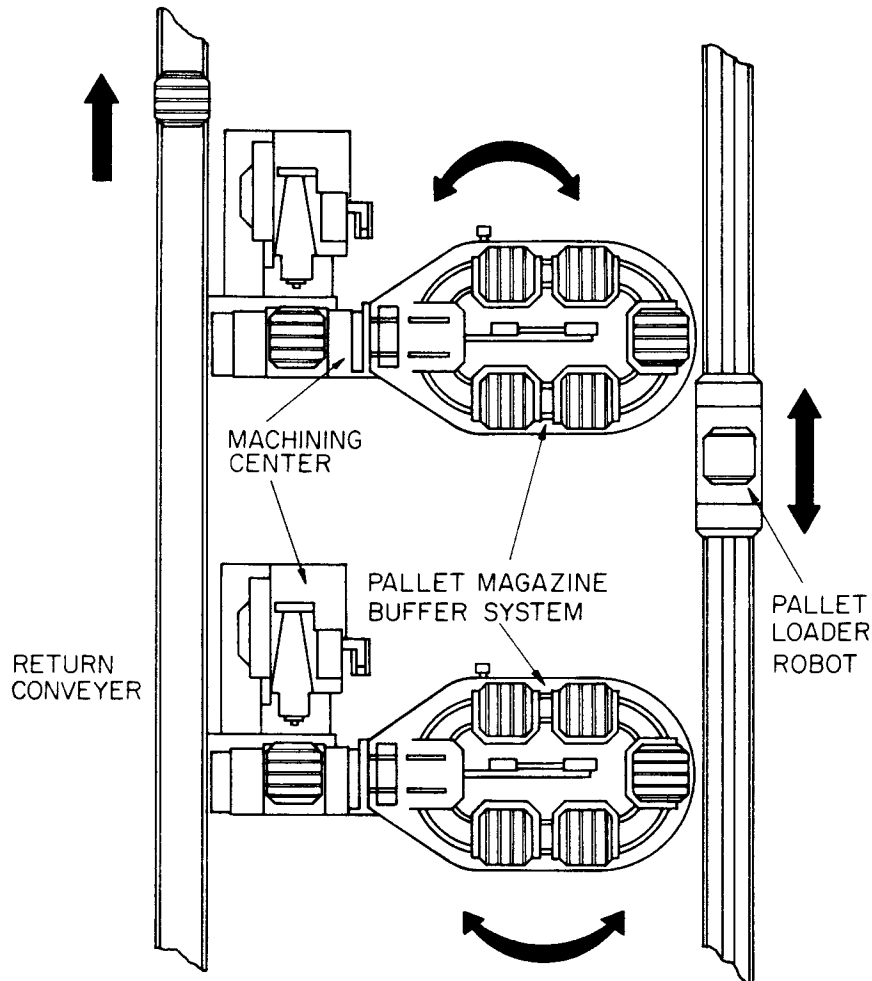


Fig. 1. Layout of two work stations with pallet buffers.

centralized material handling station, to load a new part, to route it to its next designation station, and finally to unload it there. The manufacturing time at work station k is assumed to be *exponentially distributed*, with a mean of $1/\lambda_k$, $1 \leq k \leq M$. In reality, however, manufacturing times are not always exponentially distributed. There are several theoretical and empirical studies which indicate that this assumption may not lead to a significant loss in accuracy [2,20,21,25,35]. At the cost of a large increase of the number of states, the manufacturing times could be taken to be Erlang or phase type.

The work stations take raw parts, one at a time, from the local input buffers as long as parts are available. A shortage penalty of C_k , $1 \leq k \leq M$, dollars per hour is incurred by station k when it becomes idle because of the absence of parts. The production control problem is to determine which work station should be fed next in order to minimize the expected shortage penalty. Such a decision is made by the controller (e.g. an industrial microprocessor) having full knowledge of the instantaneous buffer stocks [10,25,33] whenever the MHS becomes available. Possible inter-

action between the routing scheme discussed here and the higher-level FMS controls enforcing due-dates and production mix are outlined in [19,36].

Next, the optimal routing strategy is obtained by formulating and solving the problem as a continuous-time, finite-state, infinite-horizon undiscounted semi-Markovian decision process [16].

2.2. THE STATE SPACE

Let n_k ($1 \leq k \leq M$) denote the number of parts at station k , either at the input buffer or possibly the one part being processed. The state space is then:

$$\Omega = \{ \mathbf{n} = (n_1, n_2, \dots, n_M) \mid 0 \leq n_k \leq B_k, \quad (k = 1, 2, \dots, M) \}, \quad (2)$$

and the total number of states is

$$NS = \prod_{i=1}^M (B_i + 1). \quad (3)$$

2.3. THE TRANSITION PROBABILITIES

The joint transition probability that station i , $1 \leq i \leq M$, uses up j_i of its n_i parts during the time the MHS delivers one part to station k , $1 \leq k \leq M$, is denoted by

$$TP^k(\mathbf{n}, \mathbf{j}) \text{ with } \mathbf{n} = (n_1, n_2, \dots, n_M) \text{ and } \mathbf{j} = (j_1, j_2, \dots, j_M).$$

Next, let $P_i^k(n_i, j_i, z)$ define the *conditional probability* that station i having n_i parts consumes j_i parts during the time interval z . Hence, we conclude that in general

$$TP^k(\mathbf{n}, \mathbf{j}) = \int_0^\infty \prod_{i=1}^M P_i^k(n_i, j_i, z) dG_k(z). \quad (4)$$

Station i can consume at most n_i parts during time z . This results in:

$$P_i^k(n_i, j_i, z) = \begin{cases} \frac{(\lambda_i z)^{j_i} e^{-\lambda_i z}}{j_i!} & n_i \geq 1 \\ & j_i = 0, 1, \dots, n_i - 1 \\ 1 - \sum_{m=0}^{n_i-1} \frac{(\lambda_i z)^m e^{-\lambda_i z}}{m!} & n_i \geq 1 \\ & j_i = n_i \\ \delta_{j_i 0} & n_i = 0 \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Combining the four cases of (5) and then inserting (5) into (4) and integrating over z leads to a closed form expression [24] for the transition probability in the case that $dG_k(z)$ is given by (1).

2.4. THE DECISION SCOPE

Transitions between system states may occur at any time. Each transition in this continuous-time model starts, and terminates, at a possible decision epoch. Note that the MHS can be either blocked (i.e. all M buffers are full), or unblocked:

(a) *Unblocked MHS: $\mathbf{n} \neq \mathbf{B} = (B_1, B_2, \dots, B_M)$*

The MHS has just delivered a part to the appropriate input buffer. It must now decide whether to *remain idle* until the next change in the system state occurs (denote this decision as $k = 0$), or to *route* another part k , ($k = 1, 2, \dots, M$). The MHS cannot route a part to a blocked station; hence, the work station to be served next must be selected from only those stations that have nonfull buffers, i.e. from:

$$D(\mathbf{n}) = \{k \mid n_k < B_k\}. \quad (6)$$

(b) *Blocked MHS: $\mathbf{n} = \mathbf{B}$*

Now all buffers are full and the MHS remains idle until the next decision epoch when the first buffer becomes empty. (At this next decision epoch, the MHS can either begin routing a part to that empty buffer or else decide to remain idle until an end of processing event occurs at some other work station.) The mean time the system will hold in this state is $1/\lambda(\mathbf{B})$, where:

$$\lambda(\mathbf{B}) = \sum_{i=1}^M \lambda_i. \quad (7)$$

2.5. THE MEAN TRANSITION COSTS

Let $TC(\mathbf{n}, k)$ denote the cost incurred during a single state transition. It is a function of the current state (\mathbf{n}) and of the decision taken, $k = (0, 1, 2, \dots, M)$. This cost is given by:

$$TC(\mathbf{n}, k) = \sum_{i=1}^M SC_i(n_i, k). \quad (8)$$

Computing $SC_i(n_i, k)$, note that the time t required by station i to consume all its n_i parts has the following distribution:

$$\theta(n_i, t) = \frac{(\lambda_i)^{n_i} t^{n_i-1}}{(n_i-1)!} e^{-\lambda_i t}; \quad \begin{matrix} t \geq 0 \\ n_i \geq 1 \end{matrix} \quad (9)$$

Denote by z the actual MHS travel time following a decision k , $k = 0, 1, \dots, M$. The mean starvation interval follows from the expected value of $\text{Max}[0, z - t]$. When the shortage penalty per unit time for station i is C_i , then

$$SC_i(n_i, k) = \begin{cases} C_i \int_0^\infty \int_t^\infty (z-t) \theta(n_i, t) dt dG_k(z) & n_i > 0, k > 0 \\ C_i \int_0^\infty z dG_k(z) & n_i = 0, k > 0 \\ C_i/\lambda(n) & n_i = 0, k = 0 \\ 0 & n_i > 0, k = 0. \end{cases} \quad (10)$$

In this model, we define:

$$\begin{aligned} \lambda(n) &= \sum_{i \in A(n)} \lambda_i \text{ (the cumulative production rate of all active stations in} \\ &\text{state } n), \text{ and} \\ A(n) &= \{i \mid 0 < n_i \leq B_i\}. \end{aligned}$$

Explicit terms for $SC_i(n_i, k)$ in the case where the MHS times have Erlang distribution are given in the appendix.

2.6. SYSTEM CONTROLS

Minimizing the expected shortage penalty cost per time unit, we define a set of functional dynamic programming [14] equations (11)–(14) as a function of the relative state values $v(n)$ and of the long run expected cost g . The control parameter is $k (= 0, 1, \dots, M)$.

When the MHS is unblocked, then:

$$\begin{aligned} v(n) = \min \left[\min_{k \in D(n)} \left\{ TC(n, k) - g/\mu_k + \sum_{0 \leq j \leq n} TP^k(n, j) v(n-j+e^k) \right\}, \right. \\ \left. \left\{ TC(n, 0) - g/\lambda(n) + \sum_{i \in A(n)} v(n-e^i)(\lambda_i/\lambda(n)) \right\} \right] \quad (11) \\ n \in \Omega, n \neq \mathbf{0} \text{ or } B, \end{aligned}$$

The *first* set of terms on the right-hand side of (11) is the relative value of state \mathbf{n} if a decision is made to route a part to station $k \in D(\mathbf{n})$; the mean holding time interval following this decision is $1/\mu_k$. During that time interval (which can extend over several changes in the system state), \mathbf{n} drops by \mathbf{j} and n_k is incremented by *one*. The *second* set of terms on the right-hand side of (11) is the relative value of state \mathbf{n} following a decision to keep the MHS idle until the next system state (i.e. for $1/\lambda(\mathbf{n})$ time units, $\mathbf{n} \neq \mathbf{0}$).

The relative value for the blocked case is:

$$v(\mathbf{B}) = -g/\lambda(\mathbf{B}) + \sum_{i=1}^M \{v(\mathbf{B} - \mathbf{e}^i) [\lambda_i/\lambda(\mathbf{B})]\} \quad \mathbf{n} = \mathbf{B}. \quad (12)$$

Note that the mean holding time in the *blocked* case is $1/\lambda(\mathbf{B})$, $\mathbf{n} = \mathbf{B}$, and since all M stations are active, no starvation penalty is then incurred (eq. (12)). The probability that station i will be the first to become unblocked is $\lambda_i/\lambda(\mathbf{B})$.

Finally, eqs. (13) and (14) are used to fix the arbitrary additive component of the relative value vector and to determine the desired routing (k) at $\mathbf{n} = \mathbf{0}$. Clearly, it is assumed that the MHS *can not* remain idle at that state:

$$v(\mathbf{0}) = \min_{k=1, \dots, M} \{TC(\mathbf{0}, k) - g/\mu_k + v(\mathbf{e}^k)\} \quad \mathbf{n} = \mathbf{0} \quad (13)$$

$$v(\mathbf{0}) = 0. \quad (14)$$

Solving these functional equations yields: the *optimal routing policy* for the MHS (k as a function of \mathbf{n}), the expected *cost per unit time* ($=g$) following the optimal policy and the *relative value* of each state $v(\mathbf{n})$ with respect to $v(\mathbf{0})$ – where no parts are present in the local buffers. The computation of several key performance measures is detailed in sect. 3.

2.7. THE MODEL SOLUTION

Equations (11)–(14) possess a unique solution [22,23], since under any policy it is possible to reach state $\mathbf{n} = \mathbf{0}$ with positive probability (when the MHS becomes temporarily sluggish). Consequently, every transition probability matrix is unichain, with $\mathbf{0}$ as the regeneration point, plus possibly some transient states as well. This is sufficient to ensure [23] that a unique solution exists to (11)–(14). These $NS + 1$ equations are solved by augmenting the value iteration scheme of [22] with a stepsize

$$\tau = \min \left[1/\lambda(\mathbf{B}), \min_{1 \leq k \leq M} 1/\mu_k \right], \quad (15)$$

and an initial guess $v(\mathbf{n}) = -|\mathbf{n}|$ for all \mathbf{n} . The termination criteria was to exit when g could be estimated within $\pm 0.1\%$.

A "semi-Markovian solver" algorithm was developed for this problem, along with a computer program that generates the optimal routing policy and computes the desired performance measures. Memory requirements are a linear function of NS, while the observed time complexity is $O(NS^2)$. Analysis of a problem having one hundred states requires, on the average, 20 to 40 iterations and a few CPU minutes on a CDC 855/170 computer with the NOS 2.3 operating system.

3. Performance measures

Following the solution of the functional model equations and the determination of the optimal (or the desired) control policy – denoted as $k^*(\mathbf{n})$ – it is possible to compute the FMS's principal performance measures. These measures are evaluated by varying the transition costs accordingly.

3.1. THE PRODUCTION RATES

Let r_t , $t = 1, 2, \dots, M$, denote the production rate of station t following $k^*(\mathbf{n})$. It is also the number of parts transported per unit of time. Define a new semi-Markovian process producing a unit "reward" whenever the MHS routes one part of type t , i.e.:

$$q(\mathbf{n}) = \begin{cases} 1 & \text{if } k^*(\mathbf{n}) = t \quad (t = 1, 2, \dots, M), \mathbf{n} \in \Omega, k^*(\mathbf{n}) \in 1, 2, \dots, M \\ & \text{[MHS delivers part type } k^*(\mathbf{n})] \\ 0 & \text{otherwise.} \end{cases} \quad (16)$$

The associated equations for computing the throughputs of type t are.

$$v(\mathbf{n}) = q(\mathbf{n}) - r_t/\mu_k^*(\mathbf{n}) + \sum_{0 \leq j \leq \mathbf{n}} \text{TP}^{k^*(\mathbf{n})}(\mathbf{n}, j) v(\mathbf{n} - j + e^{k^*(\mathbf{n})}) \quad (17)$$

$$k^*(\mathbf{n}) = 1, \dots, M, \mathbf{n} \in \Omega.$$

$$\text{[MHS delivers part type } k^*(\mathbf{n})]$$

$$v(\mathbf{n}) = q(\mathbf{n}) - r_t/\lambda(\mathbf{n}) + \sum_{i \in A(\mathbf{n})} \frac{\lambda_i}{\lambda(\mathbf{n})} v(\mathbf{n} - e^i) \quad (18)$$

$$k^*(\mathbf{n}) = 0, \mathbf{n} \in \Omega, \mathbf{n} \neq \mathbf{0}$$

$$\text{(MHS remains idle)}$$

$$v(\mathbf{0}) = 0. \quad (19)$$

The value scheme of [22] is adapted to solve these equations. The long-run expected reward per unit time is r_t , the throughput of station t .

3.2. THE UTILIZATION OF THE WORK STATIONS

Given r_t , the utilization of station t is

$$U_t = r_t / \lambda_t \quad t = 1, 2, \dots, M. \quad (20)$$

3.3. THE UTILIZATION OF THE MHS

Since r_t / μ_t is the fraction of the MHS time it spends delivering parts of type t , the MHS's utilization is

$$UH = \sum_{t=1}^M r_t / \mu_t. \quad (21)$$

Here, $(1 - UH)$ is the fraction of time during which the MHS is idle (i.e. the system is either in state $\mathbf{n} = \mathbf{B}$ or in state $\mathbf{n} \neq \mathbf{B}$ and $k^*(\mathbf{n}) = 0$).

3.4. THE NUMBER OF STARVATION PERIODS (AT THE STATIONS)

$SP(i | n_i, k)$, ($i = 1, 2, \dots, M$) denotes the probability that station i will *start* starving, during the time interval following decision k , ($k = 0, 1, 2, \dots, M$), given that this time interval began with $n_i \geq 1$ parts at station i . Let t denote the time for station i to consume *all* its $n_i \geq 1$ parts and let z denote the MHS delivery time. The probability density function of t is given by (9). First, consider the starvation probability for $k = 1, 2, \dots, M$ and for $n_i \geq 1$. In this case, we get that:

$$\begin{aligned} SP(i | n_i, k) &= \Pr [t \leq z] \\ &= \int_0^{\infty} \theta(n_i, t) dt \int_t^{\infty} dG_k(z) \quad k \geq 1, n_i \geq 1, t > 0. \end{aligned} \quad (22)$$

In the case that $dG_k(z)$ is given by (1), then the following closed form expression is derived ([24]):

$$\begin{aligned}
 SP(i|n_i, k) &= \lambda_i L_k \mu_k \sum_{j=0}^{L-1} \left(\frac{L_k \mu_k}{\lambda_i} \right)^j \frac{(j+n_i-1)!}{(n_i-1)!j!} (1 + L_k \mu_k / \lambda_i)^{-(j+n_i)} \\
 &= \frac{\lambda_i L_k \mu_k}{(1 + L_k \mu_k / \mu_i)^{n_i}} \sum_{j=0}^{L_k-1} \binom{j+n_i-1}{j} \left[\frac{L_k \mu_k}{L_k \mu_k + \lambda_i} \right]^j
 \end{aligned} \tag{23}$$

$$L_k \geq 1, n_i \geq 1, k = 1, 2, \dots, M.$$

If $n_i = 0$ at the *beginning* of the interval (i.e. starvation commenced at some previous time interval), we obtain:

$$SP(i|n_i, k) = 0 \quad k \geq 0, n_i = 0. \tag{24}$$

If the decision is to *halt* the MHS activities (i.e. $k = 0$) until the next part is manufactured, then

$$SP(i|n_i, k) = \lambda_i / \lambda(\mathbf{n}) \quad \text{since } k = 0, n_i = 1. \tag{25}$$

In all other cases (i.e. $k = 0$ and $n_i \geq 1$), this probability is equal to zero.

We get, therefore, the following scheme for computing the values of $SP(i|n_i, k)$:

$$SP(i|n_i, k) = \begin{cases} \text{See (22) or (23)} & k = 1, 2, \dots, M \quad n_i > 1, \\ \frac{\lambda_i}{\lambda(\mathbf{n})} & k = 0, n_i = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

To compute the expected number of the starvation periods per unit time at station i , NSP_i , $i = 1, \dots, M$, define a semi-Markovian process similar to (17)–(19) for station i , but here the gain rate is NSP_i and the mean-transition reward is set to $SP(i|n_i, k)$. The desired values of NSP_i are arrived at by solving the resulting functional equations.

3.5. THE EXPECTED DURATION OF EACH STARVATION PERIOD (AT THE STATIONS)

The fraction of time that station t is idle is equal to $(1 - U_t)$. The expected duration of the starvation periods of station t is

$$SD_t = (1 - U_t) / NSP_t. \tag{27}$$

3.6. THE EXPECTED DURATION OF EACH BLOCKED PERIOD (AT THE MHS)

When all the local buffers are full, the MHS is blocked. As noted earlier, the expected duration of each blockage period at the MHS is

$$BD = 1/\lambda(\mathbf{B}). \quad (28)$$

3.7. THE MEAN BUFFER OCCUPANCIES (AT DECISION EPOCHS)

The mean duration of the time interval between two consecutive decision epochs following any given policy is denoted by TI. To compute TI, define a semi-Markovian process similar to (17)–(19) with a one-transition reward $q(\mathbf{n})$ equal to *one*. The computed gain rate then is $1/\text{TI}$.

Next, let MBO_t denote the mean buffer occupancy at station i , $i = 1, 2, \dots, M$, at the FMS's decision epoch. For station t , we define a semi-Markovian process similar to the one mentioned above, but with the following one-transition reward:

$$q(\mathbf{n}) = n_t \quad (t = 1, 2, \dots, M). \quad (29)$$

The computed gain rate of this process is the mean buffer occupancy time interval, denoted as MBOT_t . This leads directly to the final result:

$$\text{MBO}_t = \text{MBOT}_t \text{ TI}. \quad (30)$$

This performance measure provides an indicator to the state space structure, as seen by the MHS following a given policy. It shows the disparity in the n_i values at the decision epochs.

4. Numerical illustrations

First, consider the interaction between the structure of the optimal policy and the allocation of buffer spaces within the FMS. This interaction is illustrated in two examples denoted as Instances 1 and 2 – both detailed in table 1. The values of B_1 and B_2 were pushed there to the extremes in order to clearly exhibit some characteristic features of the optimal routing policies.

Figures 2 and 3 depict the optimal decision boundaries for instances 1 and 2, respectively. They show that the boundary curves between the two decision regions are non-decreasing as a function of n_1 and n_2 . This means that if the optimal decision for a given state $\{\mathbf{n} = \tilde{n}_1, \tilde{n}_2\}$ is, say, $k = 2$, then it will be the same k for all unblocked states having $\tilde{n}_1 \leq n_1 \leq B_1$ and $n_2 = \tilde{n}_2$, or vice versa for $k = 1$. Figure 2 demonstrates that the optimal policy does not utilize all the available buffer spaces: the last two ones in work station 1 belong to the zero decision boundary, since $k(8, 1) = k(9, 1) = 0$.

Table 1
FMS parameters for Instances 1 and 2

Parameter	Instance			
	1		2	
Buffer spaces (B_1, B_2)	10	1	7	4
Production rates: (λ_1, λ_2) (parts/hour)	20	100	20	100
MHS capacity: (μ_1, μ_2) (parts/hour)	100	100	100	100
Starvation penalties: (C_1, C_2) (\$/hour-idle)	90	90	90	90

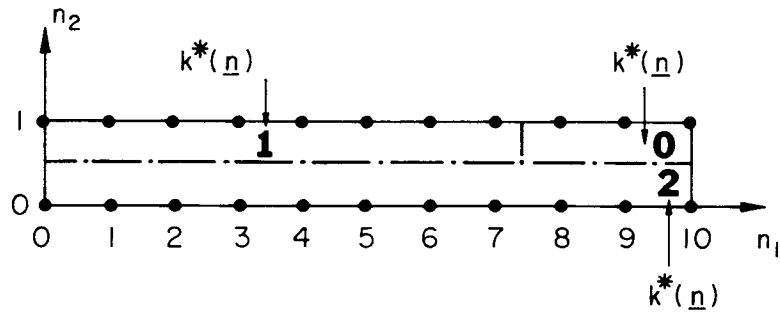


Fig. 2. Optimal routing strategy for instance 1.

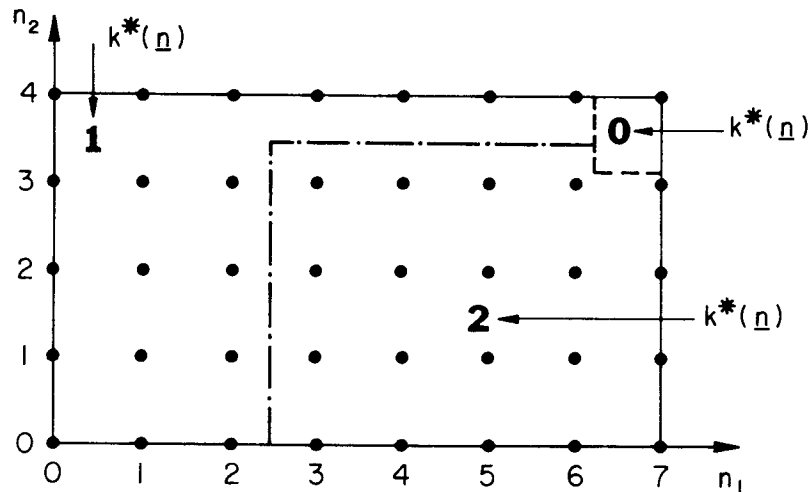


Fig. 3. Optimal routing strategy for instance 2.

Table 2
A comparison of two FMS configurations: instance 1 versus instance 2

Performance measures	i	Instance 1	Instance 2
r_i	1	20.0	19.90
	2	44.95	37.57
U_i	1	1.0	0.9953
	2	0.4495	0.3757
NSP $_i$	1	0	31.33
	2	36.94	20.25
SD $_i$	1	0	0.00015
	2	0.0149	0.01305
MBO $_i$	1	9.72	4.91
	2	0.58	2.385
BD		1/120	1/120
UH		0.6498	0.9347
g		49.54	24.21

This means that in these states the MHS is kept idle, apparently sparing capacity to route additional parts to work station 2.

Table 2 shows that significant performance improvements are gained by merely changing the internal allocation of the eleven buffer spaces between the two work stations. Performance measures presented in table 2 are: work stations throughputs (r_i), work stations utilization (U_i), number of starvation periods at each station (NSP $_i$), expected duration of each starvation period at a given station (SD $_i$), mean buffer occupancies at the decision epochs (MBO $_i$), expected duration of each blockage period at the MHS (BD), MHS utilization (UH), and the mean starvation penalty per time unit (g).

Altogether, these results seem to contra-indicate the arbitrary allocation of buffer spaces to work stations in FMSs. This important design parameter should be considered in the light of the operational objectives and the routing policy to be applied at that facility.

Next, consider three additional data sets having the same FMS parameters but with distinct MHS routing times variability. In these cases assume: $M = 2$, $B_1 = B_2 = 3$, $\lambda_1 = 50$, $\lambda_2 = 100$, $C_1 = C_2 = 90$, and $\mu_1 = \mu_2 = 100$. Also assume that

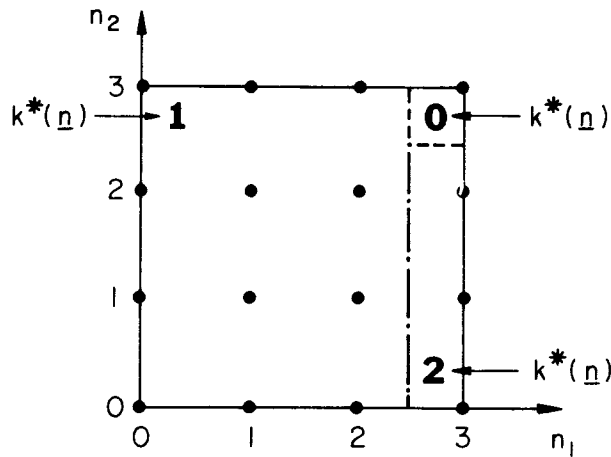


Fig. 4. Optimal routing strategy for instances 3, 4, and 5.

Table 3
FMS performance with Erlang distributed routing times

Performance measures	i	Instance 3 ($L_1 = L_2 = 1$)	Instance 4 ($L_1 = L_2 = 2$)	Instance 5 ($L_1 = L_2 = 5$)
r_i	1	44.40	45.95	46.96
	2	48.76	49.39	49.82
U_i	1	0.8881	0.9190	0.9393
	2	0.4877	0.4939	0.4983
NSP_i	1	22.11	18.75	16.02
	2	18.76	21.11	23.08
SD_i	1	0.00506	0.0043	0.0038
	2	0.0273	0.0240	0.02170
MBO_i	1	2.41	2.3969	2.3917
	2	1.31	1.1891	1.0874
BD		1/150	1/150	1/150
UH		0.9317	0.9534	0.9679
g		56.19	52.84	50.62

$L_1 = L_2 = 1$ for instance 3, $L_1 = L_2 = 2$ for instance 4, and $L_1 = L_2 = 5$ for instance 5. The computed optimal policies for these three instances were *identical* (see fig. 4). This finding illustrates the robustness of the optimal policy in this case. Comparison of the performance measures in table 3 reveals a slight improvement in the FMS's performance as L_k increases. This is an expected result, attributable to the apparent reduction in the routing time variability.

Table 4
FMS performance with optimal and heuristic routing strategies

Performance measures	i	Instance 3		Instance 6	
		OPTIMAL	SQ	OPTIMAL	SQ
r_i	1	44.40	35.14	49.16	36.36
	2	48.76	60.23	49.38	63.25
U_i	1	0.8881	0.7028	0.9832	0.7272
	2	0.4877	0.6024	0.4938	0.6325
SD_i	1	0.0506	0.0080	0.0027	0.0078
	2	0.0273	0.0167	0.0320	0.0167
MBO_i	1	2.41	1.5322	4.952	1.814
	2	1.31	1.5942	1.617	1.890
BD		1/150	1/150	1/150	1/150
UH		0.9317	0.9538	0.9854	0.9960
g		56.19	62.53	47.06	57.64

Finally, in table 4 the performance of the optimal routing strategy is compared with that of the shortest queue (SQ) rule [9,34]. Two instances are considered here. The first is instance 3 described earlier, and the second is instance 6, which is similar to instance 3 except that $B_1 = B_2 = 6$.

Figure 5 depicts the routing strategies for instances 3 and 6. When using SQ, the fastest station (i.e. 2) is selected in a case of a tie.

The starvation penalty (g) generated by the SQ policy is 11.3% and 22.5% greater than the optimal values in instances 3 and 6, respectively. In both cases, the SQ policy results in higher MHS utilization which, in turn, leads to delivery delays and to increased stations starvation. Since SQ attempts at balancing the queue lengths (n_i) in front of every station, it was therefore anticipated that under this policy the mean buffer occupancies at the decision epochs (MBO_i) will be similar.

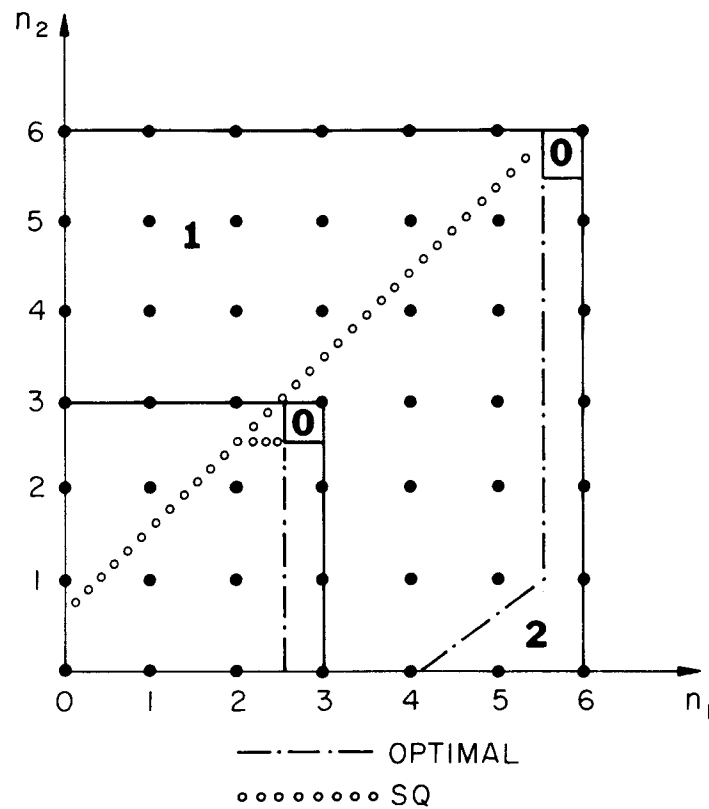


Fig. 5. Routing strategies for instances 3 and 6.

5. Conclusions and industrial implementations

This paper presents a new tractable analytical formulation for job routing in FMSs. The computed performance measures include, for example, work stations and MHS utilization, throughputs, mean buffer occupancies (at the decision epoch), the frequency, and the expected duration of each starvation period at the stations. Since the routing system is modeled as a semi-Markovian decision process, various distribution functions of the MHS can be accommodated. The average computational effort involved in using these models is a quadratic function of the number of states. Numerical experimentations with problems having several hundred states indicate that solving the functional equations of the suggested model makes only modest computational demands on core memory and CPU time.

Implementation of the models discussed here may be in the area of real-time on-line operational control of FMSs, where stable but reactive response is required in highly uncertain environments. Another potential industrial application of these models is in the area of stochastic performance evaluation of FMSs during the planning and the design phases of these systems [21,26,29]. One can, for instance, assess the impact of varying the number of work stations, the allocation of buffer spaces, or the

MHS capacity on the expected work stations utilization and throughputs [18]. Finally, these formulations can be used in conjunction with the optimal design model recently developed by Yao and Shanthikumar [30]. Their model determines the desired allocation of a raw parts input stream among a finite set of distinct manufacturing cells. Unlike this paper, it is assumed in [38] that parts must exit the system (to an overflow facility) if all the machines at their designated cell are busy. This loss probability is kept below a given limit.

Appendix

THE ERLANG MEAN TRANSITION COSTS

Following (1), (9) and (10), we get for $n_i > 0$, $k > 0$ that

$$\begin{aligned} SC_i(n_i, k) &= C_i \int_0^{\infty} \frac{\lambda_i^{n_i}}{(n_i - 1)!} t^{n_i - 1} e^{-\lambda_i t} dt \int_t^{\infty} \frac{(L_k \mu_k)^{L_k}}{(L_k - 1)!} z^{L_k - 1} e^{-L_k \mu_k z} (z - t) dz. \quad (\text{A.1}) \end{aligned}$$

Define the following auxiliary function

$$\begin{aligned} ux(t^{n_i}, z^{L_k}) &= \int_0^{\infty} t^{n_i} e^{-\lambda_i t} dt \int_t^{\infty} z^{L_k} e^{-L_k \mu_k z} dz \\ &= \frac{(n_i + L_k)!}{L_k \mu_k (\lambda_i + L_k \mu_k)^{n_i + L_k + 1}} + \frac{1}{\mu_k} \int_0^{\infty} t^{n_i} e^{-\lambda_i t} dt \int_t^{\infty} z^{L_k - 1} e^{-(L_k \mu_k z)} dz. \quad (\text{A.2}) \end{aligned}$$

For $L_k = 0$, we obtain the seed of the integrals:

$$\begin{aligned} ux(t^{n_i}, 0) &= \int_0^{\infty} t^{n_i} e^{-\lambda_i t} dt \int_t^{\infty} e^{-\mu_k z} dz \\ &= \frac{1}{\mu_k} \frac{n!}{(\lambda_i + \mu_k)^{n_i + 1}}. \quad (\text{A.3}) \end{aligned}$$

Using (A.2), we perform the integration which finally leads to

$$SC_i(n_i, k) = \frac{\lambda_i^{n_i} (L_k \mu_k)^{L_k} C_i}{(n_i - 1)! (L_i - 1)!} \cdot [ux(t^{n_i-1}, z^{L_k}) - ux(t^{n_i}, z^{L_k-1})],$$

$$1 \leq i \leq M, L_k = 1, 2, \dots, 1 \leq n_i \leq B_i. \quad (\text{A.4})$$

If $n_i = 0$ and $k > 0$, then

$$SC_i(0, k) = C_i / \mu_k, \quad 1 \leq i \leq M, L_k = 1, 2, \dots. \quad (\text{A.5})$$

Acknowledgements

The author would like to thank Professor P.J. Schweitzer of the W.E. Simon Graduate School of Business Administration at the University of Rochester, Professor S.B. Gershwin of the Laboratory for Information and Decision Systems at MIT, and Professor D.D. Yao of Harvard University for useful discussions. The author is also grateful to the associate editor and two anonymous referees for their suggestions and comments. Partial support for this research has been provided by the IBM Program for Support for the Education in the Management of Information Systems.

References

- [1] A. Arbel and A. Seidmann, Performance evaluation of flexible manufacturing systems, *IEEE Trans. on Systems, Man and Cybernetics* 14, 4(1984)606.
- [2] M.M. Barash, Computerized manufacturing systems for discrete products, in: *The Handbook of Industrial Engineering*, ed. G. Salvendy (Wiley, New York, 1981) Ch. VII-9.
- [3] J.A. Buzacott and D.D. Yao, On queueing network models of flexible manufacturing systems, *Queueing Systems* 1, 1(1986)5.
- [4] J.B. Cavaille and D. Dubois, Heuristic methods based on mean value analysis for flexible manufacturing systems performance evaluation, *Proc. IEEE Conf. on Decision and Control*, Orlando, Florida (1982).
- [5] C. Derman, On sequential decisions and Markov chains, *Manage. Sci.* 9, 1(1962)16.
- [6] C. Dupont-Gatelman, A survey of flexible manufacturing systems, *J. Manufacturing Systems* 1, 1(1982)1.
- [7] C. Dupont-Gatelman, A survey of analytical and simulation models for the design and control of flexible manufacturing systems, *CIRP Annals, Manufacturing Systems* 4, 2(1983) 157.
- [8] W. Eversheim and P. Hermann, Recent trends in flexible automated manufacturing, *J. Manufacturing Systems* 1, 2(1982)139.
- [9] G.T. Foschini, On heavy traffic diffusion analysis and dynamic routing in packet switched networks, in: *Computer Performance*, ed. K.M. Chandy and M. Reiser (North-Holland, New York, 1977)499–513.
- [10] S.B. Gershwin, R.R. Hildebrandt, R. Suri and S.K. Mitter, A control theorist's perspective on recent trends in manufacturing systems, *Proc. IEEE Conf. on Decision and Control*, Las Vegas, Nevada (1984).

- [11] E.H. Hahne, Dynamic routing in an unreliable manufacturing network with limited storage, Report No. LIDS-TH-1063, Laboratory for Information and Decision Systems, MIT, Cambridge, MA (1981).
- [12] J. Hatvany (ed.), *World Survey on CAM* (Butterworths, Kent, U.K., 1983).
- [13] R.R. Hildebrant, Scheduling flexible machining systems using mean value analysis, *Proc. IEEE Conf. on Decision and Control*, Albuquerque, New Mexico (1980).
- [14] R.A. Howard, *Dynamic Probabilistic Systems* (Wiley, New York, 1971) Vols. I, II.
- [15] G.K. Hutchinson, Flexible manufacturing systems in the United States, Management Research Center, University of Wisconsin-Milwaukee (1979).
- [16] W.J. Jewell, Markov-renewal programming: I and II, *Oper. Res.* 11(1963)938; 949.
- [17] J. Kimemia and S.B. Gershwin, Multi-commodity network flow optimization in flexible manufacturing systems, Report No. ESL-FR-2, Laboratory for Information and Decision Systems, MIT, Cambridge, MA (1980).
- [18] A. Kusiak, Material handling in flexible manufacturing systems, Working Paper 06/85, Department of Industrial Engineering, Technical University of Nova Scotia (1985).
- [19] T.E. Morton, Patriarch: Intelligent operations management for manufacturing, Working Paper, Carnegie Mellon University (1985).
- [20] G.J. Olsder and R. Suri, Time-optimal control of parts-routing in a manufacturing system with failure-prone machines, *Proc. 19th IEEE Conf. on Decision and Control* (1980).
- [21] K. Rathmill, N. Greenwood and M. Houshmand, Computer simulation of FMS, *Proc. 2nd Int. Conf. on FMS*, London, U.K. (1983) pp. 251 – 280.
- [22] P.J. Schweitzer, Iterative solution of the functional equations of undiscounted Markov renewal programming, *J. Math. Analysis and Applications* 34, 3(1971)495.
- [23] P.J. Schweitzer, On the existence of relative values for undiscounted Markovian decision processes with a scalar gain rate, *J. Math. Analysis and Applications* 104, 1(1984)67.
- [24] A. Seidmann, A stochastic dynamic programming approach to part routing in flexible manufacturing systems, Working Paper, W.E. Simon Graduate School of Business Administration, University of Rochester, Rochester, NY (1987).
- [25] A. Seidmann and P.J. Schweitzer, Part selection policy for a flexible manufacturing cell feeding several production lines, *IIE Trans.* 16, 4(1984)355.
- [26] A. Seidmann, P.J. Schweitzer and S. Shalev-Oren, Computerized closed queueing network models of flexible manufacturing systems: A comparative evaluation, *Large Scale Systems*, to appear.
- [27] S. Shalev-Oren, A. Seidmann and P.J. Schweitzer, Analysis of flexible manufacturing systems with priority scheduling: PMVA, *Ann. Oper. Res.* 3(1985)115.
- [28] J.J. Solberg, A mathematical model of computerized manufacturing systems, *Proc. 4th Int. Conf. on Production Research*, Tokyo, Japan (1977).
- [29] J.J. Solberg and S.Y. Nof, Analysis of flow control in alternative manufacturing configurations, *J. Dynamic Systems, Measurements and Control* 102, 9(1980)141.
- [30] K.E. Stecke, Design, planning, scheduling and control problems of flexible manufacturing systems, *Ann. Oper. Res.* 3(1985)3.
- [31] K.E. Stecke and J.J. Solberg, The optimality of unbalancing both workloads and machine group sizes in closed queueing networks of multiserver queues, *Oper. Res.* 33, 4(1985)882.
- [32] R. Suri, New techniques for modelling and control of flexible automated manufacturing systems, *IFAC Conference*, Kyoto, Japan (1981).
- [33] R. Suri and R.R. Hildebrant, Modelling flexible manufacturing systems using mean-value analysis, *J. Manufacturing Systems* 1(1984)3.
- [34] D. Towsley, Queueing network models with state-dependent routing, *J. Ass. Comp. Mach.* 27, 3(1979)323.

- [35] D.D. Yao and J.A. Buzacott, Modelling the performance of flexible manufacturing systems, *Int. J. Prod. Res.* 24, 6(1985)945.
- [36] D.D. Yao and J.A. Buzacott, Modelling a class of state-dependent routing in flexible manufacturing systems, *Ann. Oper. Res.* 3(1985)153.
- [37] D.D. Yao and J.G. Shanthikumar, The optimal input rates to a system of manufacturing cells, *INFOR* 25, 7(1987)57.
- [38] D.D. Yao and J.G. Shanthikumar, Optimal server allocation in a system of manufacturing cells, submitted for publication (1987).