DYNAMIC SCHEDULING AND ROUTING FOR FLEXIBLE MANUFACTURING SYSTEMS THAT HAVE UNRELIABLE MACHINES

Oded Z. Maimon and Stanley B. Gershwin

Abstract

This paper presents a method for real-time scheduling and routing of material in a flexible manufacturing system (FMS). It extends the earlier scheduling work of Kimemia and Gershvin. The FMS model includes machines that fail at random times and stay down for random lengths of time. The new element is the capability of different machines to perform some of the same operations. The times that different machines require to perform the same operation may differ. This paper includes a model, its analysis, a real-time algorithm, and examples.
Acknowledgements


Author Information


Copyright (c) 1987, MIT. Memos in this series are for use by MIT and are not considered to be published merely by virtue of appearing in this series. This copy is for private circulation only and may not be further copied or distributed, except for government purposes, if the paper acknowledges U. S. Government sponsorship. References to this work should be either to the published version, if any, or in the form "private communication." For information about the ideas expressed herein, contact the author directly. For information about this series, contact Microsystems Research Center, Room 39-321, MIT, Cambridge, MA 02139; (617) 253-8138.
Dynamic Scheduling and Routing For Flexible Manufacturing Systems

That Have Unreliable Machines

by

Oded Z. Maimon

and

Stanley B. Gershwin

† Computer Integrated Manufacturing
Digital Equipment Corporation
Maynard, Massachusetts 07154

ABSTRACT

This paper presents a method for real-time scheduling and routing of material in a Flexible Manufacturing System (FMS). It extends the earlier scheduling work of Kimemia and Gershwin. The FMS model includes machines that fail at random times and stay down for random lengths of time. The new element is the capability of different machines to perform some of the same operations. The times that different machines require to perform the same operation may differ. This paper includes a model, its analysis, a real-time algorithm, and examples.

1. INTRODUCTION

Purpose

The purpose of this paper is to develop an algorithm to calculate real-time loading and routing decisions for a Flexible Manufacturing System (FMS). An algorithm for calculating loading decisions for such systems has been described in earlier papers (Kimemia and Gershwin, 1983; Gershwin, Akella, and Choong, 1985; Akella, Gershwin, and Choong, 1985). An algorithm for routing decisions is described in Maimon and Choong (1985). Here, routing and loading are calculated together.

As in the earlier papers, the problem is to decide which part should be dispatched next into a set of machines. These machines are capable of performing work on a set of different part types with no time lost for setting up. Decisions are made in response to disruptions of the operation of the system caused by machine failures, and according to the surplus or backlog for each part type. Whenever a machine changes state (i.e., fails or is repaired), a new schedule and a new routing scheme is calculated via a feedback law.

ments (e.g., tolerance better than 0.005") and are adaptable (programmable) so that they can handle different types of odd components.

As a result, different operations (insertions of odd component types) can be performed by different robots, but the amount of time required for a given operation depends on the speed of the robot that performs it. Also, each robot has different configurations (e.g., tools) and inherent capabilities (e.g., accuracy and repeatability), which results in different subsets of operations that each robot can handle (with nonempty intersections among those subsets).

As a consequence, not only does the input rate of part types into the system have to be determined, but also the decision of where to send each part for each operation (among the possible alternative robots) has to be made.

Such systems are usually justified economically only if the production volume is quite high (e.g., hundreds of thousands of components inserted per year) and the variety is high. Because of their flexibility, they are expected to meet demands that vary in the short term and that require high utilization. The work presented here aims to improve system performance (e.g., to lead to higher throughput and reduced WIP while meeting production demands).

Another type of manufacturing system is comprised of conventional and advanced machining centers. The latter are capable of performing different operations, with varied capabilities. For example, some machining centers can do drilling and milling operations that otherwise require two different conventional machines. Also, there are 3- and 5-axis machining centers. The latter can do more operations than the former without changing the part fixtureing.

As in an electronic insertion system, the scheduling and routing problem in a system of several machining centers is not only to decide on the input flow rate of each part type, but also where each operation should be done among alternative machines with different capabilities. Literature Survey

In this paper we present a method that considers, at the same time, two functions -- short-term scheduling and routing -- based on a global view of the system. Many references consider just one of these functions. For example, Whitt (1986) presents a method which can be used just for the local routing decisions. Although his paper develops generic queueing methodology, we use his results to show an example of local routing considerations.

By local routing decisions we refer to a situation by which a customer (or a part) has to join one of several queues. These queues represent, for example, the input buffers to workstations. The alternative queues are those of the alternative workstations that can perform the next operation on a particular part. An example is the location decision made only when taking a global view of the system.

Routing is treated in papers by Hahne (1981), Tsitsiklis (1981), and Seidmann and Schweitzer (1984). Hahne and Tsitsiklis deal with only two choices and machines whose randomness is due to failure and repair. Seidmann and Schweitzer have many choices, but the randomness is due to variations in processing times. In all cases, the full system is not considered. Instead, only one decision point is considered, and decisions are made on a purely local basis.

By contrast, we consider the whole system and do not treat local conditions in detail. This suggests that a hierarchical decision policy, in which both kinds of decisions -- local and global -- are made separately, may be appropriate. The local decisions should be made in a way that is consistent with the decisions made on a global basis.

Outline of Paper

Section 2 states the problem. Section 3 contains our solution, which is based on dynamic programing. Section 4 describes some numerical examples and simulation results. Conclusions and new research directions are discussed in Section 5.

2. PROBLEM STATEMENT

Section 1 describes two situations in which short-term scheduling and routing decisions are required. In this section we represent such manufacturing systems with a mathematical model.
The input to the problem is the production requirements and process data in the form of process plans and routing sheets. They specify the operations that each part type has to go through, together with a partial precedence relationship among the operations. For each operation, a set of alternative machines, and the time for the operation at each machine, (and machine reliability) are specified.

We seek a feedback law which determines when each part should be released into the system and which route it should take when it enters. The release time and the route may be functions of the current repair state of each machine as well as the current production level of each part type.

Model

The FMS consists of \(M\) work stations, and work station \(m\) consists of \(L_m\) identically configured machines. A family of \(N\) part types is being produced. The production rate of part type \(n\) at time \(t\) is \(u_n(t)\).

Let \(d_n\) be the demand rate for type \(n\) parts. This is a rate that is specified by higher level decision-makers in the decision hierarchy. We assume here that it is constant over the time interval of interest. The model is unchanged if it is deterministic but time-varying, but the computation is made more difficult. Requirements are often stated in terms of production required over some specified time interval; we convert this to demand rates.

Let \(x_n(t)\) be the surplus (if positive) or backlog (if negative) of type \(n\) parts at time \(t\). It is the difference between production and demand, and is given by

\[
\frac{dx_n}{dt} = u_n(t) - d_n.
\]

The states of the work stations are given by \(a_m(t)\). This is an integer which indicates the number of machines of work station \(m\) that are operational at time \(t\). The vector \(a\) is assumed to be the state of a continuous time Markov process with rates \(\lambda\), so that

\[
\text{prob} \left[ a(t-\Delta t) = b \mid a(t) = a \right] = \lambda_{ab}. \tag{2}
\]

Recall that different work stations may be available for some operations, and that they perform them at different speeds. Routing is the decision of which work station will perform each operation. Let \(y_{mn}^k\) be the rate at which work station \(m\) performs operation \(k\) on type \(n\) parts. (Since only a few operations among all those that are possible are performed on each part type, most of these variables are 0.) The relationship between \(u_n\) and \(y_{mn}^k\) is given by

\[
u_n = \sum_k y_{mn}^k \text{ for any } k \text{ and } n, \tag{3}\]

In this section, we formulate an optimization problem whose solution is the optimal set of \(y_{mn}^k\) variables as a function of time. In Section 3, we describe a suboptimal solution.

Capacity

The rate of flow of material into the system is limited by the rate at which machines can do operations. Each operation takes a finite amount of time, and no machine can be busy more than 100% of the time. A fundamental assumption is that there is no buffering inside the system. This reduces the total work in process, but increases the need for effective routing and scheduling.

Let \(T_m\) be the amount of time that a machine in work station \(m\) requires to do operation \(k\) on a part of type \(n\). The rate at which machines of that station have to do such operations has already been defined as \(y_{mn}^k\).

During a short interval of length \(T\), the expected number of operations performed by the machines is \(y_{mn}^k T\). (It is assumed that the interval is short so that no repairs or failures take place during it.) The total amount of time that all of the machines of station \(m\) are performing operation \(k\) on part type \(n\) is \(y_{mn}^k T_m\). The expected total amount of time that the machines of station \(m\) are performing all operations on all part types is

\[
\sum_k y_{mn}^k T_m. \tag{4}
\]

The total amount of time available on all the machines of station \(m\) is \(a_m T\) if \(a_m\) machines are operational. Therefore,

\[
\sum_k y_{mn}^k T_m \leq a_m. \tag{5}
\]

To summarize, the \(y\) flow rates must satisfy the following set of equations and inequalities:

\[
y_{mn}^k \geq 0 \forall k, m, n \tag{6}
\]
\[ \sum_{k=1}^{\infty} y_k^m \leq c_m \quad \text{for every machine } m. \tag{5} \]

\[ \sum_{k=1}^{\infty} y_k^m = \sum_{k=1}^{\infty} y_k^m \quad \text{for all } k \neq k_m \]

and all part types \( n \), \tag{6}

where \( k_m \) is the name of the first operation performed on parts of type \( m \). Denote by \( \Omega(n) \) the set of all \( y \) flow rates that satisfy (4) - (6).

Note that \( \Omega(n) \) is a random set. As machines fail and are repaired the instantaneous capacity changes. The rates at which material flows into the system must change as \( \Omega(n) \) changes, as well as in anticipation of these changes.

3. Solution

Following the usual dynamic programming practice, define

\[ J(x, \alpha, t) = \min_{\alpha(\cdot)} \left[ \int_0^T g(x(s)) ds \bigg| x(t) = x, \alpha(t) = \alpha \right]. \tag{8} \]

This function satisfies the Bellman equation (Bertsekas, 1976), which takes the following form:

\[ 0 = \min_{\alpha(\cdot)} \left\{ g(x(t)) + \int_0^T \frac{2J}{2K} \left( \sum_{k=1}^{\infty} y_k^m - d_k^m \right) \right\} \]

\[ + \sum_{k=1}^{\infty} \alpha(k) J(x, 0, t) \]. \tag{9}

This equation has the following interpretation: we seek a function \( J(x, \alpha, t) \) such that the values of \( y(x, \alpha, t) \in \Omega(\alpha(t)) \) that minimize the right hand side of (9) cause that expression to be zero. This is a nonlinear partial differential equation which we cannot expect to have an analytic solution. (However, in the case of a single part type and a single machine, Akella and Kumar (1986) were able to find a closed form solution.)

If (9) has a solution, the optimal control \( y \) satisfies the following linear programming problem. Note that the cost coefficients are time-varying.

\[ \min_{y(\cdot)} \left\{ \frac{2J}{2K} \left( \sum_{k=1}^{\infty} y_k^m \right) \right\} \]

subject to

\[ y(\alpha) \]

Note that \( J \) is positive since it is the expected value of the integral of \( g \), a positive quantity. Note also that feedback law (10) minimizes

\[ \frac{dJ}{dt} = \sum_{k=1}^{\infty} \frac{2J}{2K} \left( \sum_{k=1}^{\infty} y_k^m - d_k^m \right) + \frac{2J}{2K} \]

while \( \alpha \) is constant. This is because \( y \) appears in (9) only in the same term in which it appears in (11). If \( \alpha \) remains constant long enough, and there is a \( y \) \( \in \Omega(\alpha) \) such that (11) is negative, then \( J \)
eventually reaches a minimum. We call the value of $x$ that produces this minimum the hedging point and write it as $x^*_h$. If possible, the production rate should remain at a rate that keeps $x$ at the hedging point. A positive hedging point serves as insurance for future disruptions.

After $J$ reaches this minimum, $J$ and $x$ are both constant. Therefore, at the minimum,

$$\sum y_i^a - d_a = 0 \quad (12)$$

and

$$\frac{\partial J}{\partial x} = 0 \quad (13)$$

If there is no $y \in \mathcal{A}(a)$ that satisfies (12), then $J$ cannot reach a minimum for finite $x$. That is, the production lags behind the demand requirements and $x(t)$ decreases. This is because too many machines are currently down to allow production to equal demand.

There are reasons to believe that the solution of linear programming problem (10) provides a satisfactory scheduling and routing algorithm even if an approximate $J$ function is used. This was the simulation experience reported by Gershwin, Akella, and Choong (1985) and Akella, Gershwin, and Choong (1985).

In addition, it is likely that the repair and failure processes are not actually exponential, not actually independent of the machine utilizations (assumed in Section 2), and do not have the exact $\lambda$ parameters that would be used in (9) if an exact solution could be calculated. Also, the $g$ function does not necessarily represent true costs, but rather is chosen to obtain a desired behavior. For these reasons, it would be a mistake to work very hard to get an exact $J$.

Therefore, a reasonable strategy is to select a $J$ function that has the correct qualitative properties and that is easy to calculate and work with. Such a function is positive and has a minimum at the hedging point (for every $a$ such that the demand is feasible for that $a$). Gershwin, Akella, and Choong (1985) use a quadratic function,

$$J = \frac{1}{2}A(a)x + b(a)^Tx + c(a).$$

Akella, Maimon, and Gershwin (1987) demonstrate a technique for calculating a set of values for $A(a)$, $b(a)$, and $c(a)$, from a specified $g$, for a model similar to the one presented here.

### 4. Examples

#### Example 1: Three-Machine System

Consider a three-machine system that makes two part types. Machine 1 can do operations only on Type 1; Machine 2 can only work on Type 2; and Machine 3 can do operations on both. In fact, Machine 3 can do the same operations that Machines 1 and 2 can do. Thus Type 1 parts can go to Machine 1 or Machine 3 and Type 2 parts can go to Machine 2 or Machine 3. The problem is to decide where to send each of the parts and how frequently to send them into the system.

The capacity set $\mathcal{A}(a)$ is given by:

$$r_{11}y_1^a \leq g_1 \quad (14)$$

$$r_{12}y_2^a \leq g_2 \quad (15)$$

$$r_{13}y_1^a + r_{23}y_2^a \leq g_3 \quad (16)$$

$$y_1^a, y_2^a, y_3^a, v_1^a, v_2^a \geq 0 \quad (17)$$

The production surplus and backlog dynamics are:

$$\dot{x}_1 = y_1^a + y_2^a - d_1 \quad (18)$$

$$\dot{x}_2 = y_2^a + v_2^a - d_2 \quad (19)$$

If $J(x, a, t)$ is known, then the optimal routing and scheduling policy $y$ satisfies

$$\min_{y \in \mathcal{A}(a)} \frac{\partial J}{\partial x}(y_1^a + y_2^a) + \frac{\partial J}{\partial v_1^a} (v_1^a + v_2^a) \quad (20)$$

This is a feedback control law since the constraint set is a function of $a$ and the partial derivatives are functions of $x$ and $a$. To solve this linear programming problem, several cases must be considered. Figure 1 demonstrates the various regions of $J/ix-$space that have different solutions. The regions are indicated, as well as the values of $y^a$ that are optimal in those regions. Also indi-
cated is which of the following conditions that determine the values.

\[ \frac{\partial J}{\partial x_{i}} > 0 \quad \text{(Regions I and III)} \]

\[ = y_{i}^{*} = 0, \quad y_{13}^{*} = 0 \quad \text{(A)} \]

\[ \frac{\partial J}{\partial x_{i}} > 0 \quad \text{(Regions I and II)} \]

\[ = y_{12}^{*} = 0, \quad y_{23}^{*} = 0 \quad \text{(B)} \]

\[ \frac{\partial J}{\partial x_{i}} < 0 \quad \text{(Regions II, IV, V, and VI)} \]

\[ = y_{12}^{*} = \frac{a_{3}}{r_{12}} \quad \text{(C)} \]

\[ \frac{\partial J}{\partial x_{i}} < 0 \quad \text{(Regions III, IV, V, and VI)} \]

\[ = y_{13}^{*} = \frac{a_{2}}{r_{13}} \quad \text{(D)} \]

\[ \frac{\partial J}{\partial x_{i}} < 0 \quad \text{and} \quad \frac{\partial J}{\partial x_{2}} > 0 \quad \text{(Region II)} \]

\[ = y_{13}^{*} = \frac{a_{2}}{r_{13}} \quad \text{(E)} \]

\[ \frac{\partial J}{\partial x_{i}} < 0 \quad \text{and} \quad \frac{\partial J}{\partial x_{2}} < 0 \quad \text{(Region III)} \]

\[ = y_{23}^{*} = \frac{a_{2}}{r_{23}} \quad \text{(F)} \]

If both derivatives are negative (Regions IV and V), \( y_{12}^{*} \) and \( y_{23}^{*} \) are already determined. The remaining variables, \( y_{2}^{*} (i = 1, 2) \), minimize

\[ \frac{\partial J}{\partial x_{i}} y_{2}^{*} + \frac{\partial J}{\partial x_{2}} y_{2}^{*} \]

subject to (16). The solution is

\[ \left( \frac{1}{r_{12}} \frac{\partial J}{\partial x_{i}} - \frac{1}{r_{13}} \frac{\partial J}{\partial x_{2}} \right) < 0 \quad \text{(Region V)} \]

\[ = y_{2}^{*} = \frac{a_{3}}{r_{23}} \quad \text{and} \quad y_{12}^{*} = 0 \quad \text{(G)} \]

\[ \left( \frac{1}{r_{13}} \frac{\partial J}{\partial x_{i}} - \frac{1}{r_{12}} \frac{\partial J}{\partial x_{2}} \right) > 0 \quad \text{(Region IV)} \]

\[ = y_{12}^{*} = \frac{a_{1}}{r_{12}} \quad \text{and} \quad y_{13}^{*} = 0. \quad \text{(H)} \]

In each of these regions, the control \( y_{12}^{*} \) moves the state \( x_{n} \) through the dynamics ((18) and (19)). The state moves to a boundary and then to another region. However, there is one exception. In both Regions IV and V the state moves toward the common boundary, which is given by

\[ \left( \frac{1}{r_{12}} \frac{\partial J}{\partial x_{i}} - \frac{1}{r_{13}} \frac{\partial J}{\partial x_{2}} \right) = 0 \quad \text{(Region VI). (I)} \]

If we follow rules (G) and (H), the state will move back and forth across the boundary in an unrealistic manner. This is called chattering. It occurs because the problem is singular, and a remedy is suggested by Gershin, Akella, and Choong (1985). A strategy is found which, when \( x \) reaches Region VI, keeps \( x \) in Region VI. That is, it maintains (I). It does this by determining \( y_{12}^{*} \) which minimizes (21) subject to (16) and

\[ \frac{d}{dt} \left( \frac{1}{r_{12}} \frac{\partial J}{\partial x_{i}} - \frac{1}{r_{13}} \frac{\partial J}{\partial x_{2}} \right) = 0. \quad \text{(22)} \]

This is simplified by assuming that \( J \) is quadratic:

\[ J = \frac{1}{2} x^{T} A(x) x + b(x)^{T} x + c(x). \quad \text{(23)} \]

Then

\[ \frac{\partial J}{\partial x_{i}} = A_{i} x_{i} + A_{2} x_{2} + b_{i}. \quad \text{(24)} \]

and

\[ \frac{\partial J}{\partial x_{2}} = A_{2} x_{2} + A_{3} x_{3} + b_{2}. \quad \text{(25)} \]

If

\[ \frac{\partial J}{\partial x_{1}} < 0 \]

and \( y \) is chosen so that

\[ \left( \frac{1}{r_{12}} \frac{\partial J}{\partial x_{i}} - \frac{1}{r_{13}} \frac{\partial J}{\partial x_{2}} \right) = 0, \quad \text{(26)} \]

then

\[ \frac{1}{r_{12}} (A_{2} x_{2} + A_{3} x_{3} + b_{2}) \]

\[ - \frac{1}{r_{13}} (A_{1} x_{i} + A_{2} x_{2} + b_{2}) = 0. \quad \text{(27)} \]

Since this is true for more than just one instant, its first derivative with respect to \( t \) is also 0. That is,

\[ \frac{1}{r_{12}} (A_{2} x_{2} + A_{3} x_{3} + b_{2}) \]

\[ - \frac{1}{r_{13}} (A_{1} x_{i} + A_{2} x_{2} + b_{2}) = 0, \quad \text{(28)} \]

or,

\[ \frac{1}{r_{12}} (A_{2} (y_{i}^{*} + y_{i}^{*} - d_{1}) + A_{3} (y_{i}^{*} + y_{i}^{*} - d_{1}) + b_{2}) \]

\[ - \frac{1}{r_{13}} (A_{1} (y_{i}^{*} + y_{i}^{*} - d_{1}) + A_{2} (y_{i}^{*} + y_{i}^{*} - d_{1}) + b_{2}) = 0. \quad \text{(29)} \]
From (C) and (D),

\[
\begin{align*}
\frac{1}{t_{13}} & \left[ A_{2}(\alpha_{j}/r_{j}+y_{13}-d_{j}) + A_{3}(\alpha_{j}/r_{3}+y_{13}-d_{3}) + b_{j} \right] \\
\frac{1}{t_{13}} & \left[ A_{2}(\alpha_{j}/r_{j}+y_{13}-d_{j}) + A_{3}(\alpha_{j}/r_{3}+y_{13}-d_{3}) + b_{j} \right] = 0. \quad (30)
\end{align*}
\]

Now (16) (as an equality) and (30) are two equations in two unknowns, \( y_{13} \) and \( y_{23} \). The solution is

\[
y_{13} = \frac{1}{t_{13}} \left( \frac{A_{2} - (\alpha_{j} - d_{j})}{t_{2}} - \frac{A_{3} - (\alpha_{j} - d_{j})}{t_{3}} + \beta \right)
\]

and

\[
y_{23} = \frac{1}{t_{13}} \left( \frac{A_{2} - (\alpha_{j} - d_{j})}{t_{2}} - \frac{A_{3} - (\alpha_{j} - d_{j})}{t_{3}} + \beta \right) \quad (31)
\]

After \( x \) arrives at Region VI, it stays in Region VI if \( y_{13} \) and \( y_{23} \) are given by (C) and (D) and \( y_{13} \) and \( y_{23} \) are given by (31) and (32). Chattering is avoided.

5. CONCLUSIONS

This paper presents an extension to the earlier Kimemia and Gershwin work to add a real-time routing calculation to real-time scheduling. Thus this model can be used for many more types of manufacturing systems.

Future work will include the development of local operational rules which follow the system routing decisions calculated here, and extensive simulation of various types of industries to further demonstrate the use of this work.

REFERENCES


Figure 1. Control regions in $\frac{\delta J}{\delta x}$ space.
END
Feb.
1988
DTIC