December 1987 LIDS-P-1727

MULTIPLE TIME SCALE ANALYSIS OF MANUFACTURING SYSTEMS

Adam Caromicoli [1]

Alan S. Willsky [1]

Stanley B. Gershwin [2]

Abstract

In this paper we use results on the aggregation of singularly perturbed Markov chains to analyze manufacturing systems. The basis for this analysis is the presence in the system of events and processes that occur at markedly different rates - - operations on machines, set-ups, failures, and repairs, etc. The result of the analysis is a set of models, each far simpler than the full model, describing system behavior over different time horizons. In addition, we present a new theoretical result on the computation of asymptotic rates of particular events in perturbed Markov processes, where an "event" may correspond to the occurrence of one of several transitions in the process. We may apply this result to compute effective production rates at different time scales, taking into account the occurrence of setups and failures.

[1]Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA 02139. The work of these authors was supported by the Air Force Office of Scientific Research under Grant AFOSR-88-0032, and in part by the Army Research Office under Grant DAAL03-86-K-0171. The work of the first author was also supported in part by an NSERC of Canada Fellowship.

[2]Laboratory for Manufacturing Productivity, M.I.T., Cambridge, MA 02139. The work of this author was supported by the National Science Foundation under a subcontract from Boston University on Grant DMC-8615560.

I. Introduction

In this paper we analyze a finite-state Markov chain model of a flexible manufacturing system (FMS). Such models have been used before and they allow us to perform a variety of computations and make corresponding conclusions without the additional complexity resulting from the use of a more complex model. The point of our analysis is to demonstrate the multiple time scale structure of an FMS and the resulting hierarchical computations that can be performed. Since our Markov analysis methods carry over to more general models (such as semi-Markov processes), the general nature of the conclusions and methods we describe here continue to hold for more accurate and complex FMS models.

This paper is organized as follows. In the next section, we describe a simple FMS, and we apply the methods of [1] to analyze this system. In section III, we present a new theoretical result on the asymptotic frequencies of particular events in finite-state Markov chains and again apply this to our FMS example. Our presentation is necessarily brief, and we refer the reader to [2] for a complete development.

II. Multiple Time Scale Analysis of a Simple FMS Model

Our objective in this section is to describe and analyze a Markov chain model of an FMS which results in a probabilistic evolution of the form

$$\dot{\underline{x}}^{(0)}(t) = \underline{A}^{(0)}(\epsilon)\underline{x}^{(0)}(t) \tag{2.1}$$

where $\underline{x}^{(0)}(t)$ is the vector of probabilities for the various states in the model. Here ϵ is a small parameter introduced to capture the fact that different transitions occur with rates that may differ significantly. The superscript "(0)" indicates that this is the zero*th* level in a hierarchical decomposition of our model that leads to significant

computational savings.

The hierarchical decomposition of (2.1) as developed in [1] can be described as follows. Suppose that we have our model described at the kth level as

$$\dot{\underline{x}}^{(k)}(t) = \epsilon^k \underline{A}^{(k)}(\epsilon)\underline{x}^{(k)}(t) \tag{2.2}$$

where $\underline{x}^{(k)}(t)$ is the vector of probabilities of an aggregated version of the process at the preceding level. Then, let $\underline{U}^{(k)}(0)$ denote the matrix of ergodic probabilities for $\underline{A}^{(k)}(0)$, i.e. each column of $\underline{U}^{(k)}(0)$ is the ergodic probability vector for a distinct ergodic class of $\underline{A}^{(k)}(0)$. Also, let $\underline{V}^{(k)}(\epsilon)$ be the matrix of ϵ-dependent membership matrices. That is, for the Ith ergodic class of $\underline{A}^{(k)}(0)$, and the jth state of the process associated with $\underline{A}^{(k)}(\epsilon)$, we have

$$\begin{aligned} v_{Ij}^{(k)}(\epsilon) \;=\; & \Pr(\text{Process first enters I}th\text{ ergodic class} \mid \\ & \text{process starts in state j}) \end{aligned} \tag{2.3}$$

where (2.3) is computed using $\underline{A}^{(k)}(\epsilon)$. Furthermore, let $\tilde{\underline{V}}^{(k)}(\epsilon)$ be any modification of $\underline{V}^{(k)}(\epsilon)$ such that (a) the leading order terms of each element of $\underline{V}^{(k)}(\epsilon)$ and $\tilde{\underline{V}}^{(k)}(\epsilon)$ are the same, and (b)

$$\underline{1}\,\tilde{\underline{V}}^{(k)}(\epsilon) = \underline{1} \tag{2.4}$$

i.e. the concept of "membership" is preserved. Then define

$$\underline{A}^{(k+1)}(\epsilon) = \frac{1}{\epsilon}\tilde{\underline{V}}^{(k)}(\epsilon)\underline{A}^{(k)}(\epsilon)\underline{U}^{(k)}(0) \tag{2.5}$$

so that $\underline{A}^{(k+1)}(\epsilon)$ has one state for each ergodic class of $\underline{A}^{(k)}(0)$. The main result in [1] is that

$$\begin{aligned} e^{\underline{A}^{(0)}(\epsilon)t} \;=\; & e^{\underline{A}^{(0)}(0)t} \\ & + \underline{U}^{(0)}(0)e^{\underline{A}^{(1)}(0)\epsilon t}\underline{V}^{(0)}(0) - \underline{U}^{(0)}(0)\underline{V}^{(0)}(0) \end{aligned}$$

$$
\begin{aligned}
&+\underline{U}^{(0)}(0)\underline{U}^{(1)}(0)e^{\underline{A}^{(2)}(0)\epsilon^2 t}\underline{V}^{(1)}(0)\underline{V}^{(0)} \\
&-\underline{U}^{(0)}(0)\underline{U}^{(1)}(0)\underline{V}^{(1)}(0)\underline{V}^{(0)}(0) \\
&\vdots \\
&+\underline{U}^{(0)}(0)\cdots\underline{U}^{(K-2)}(0)e^{\underline{A}^{(K-1)}(0)\epsilon^{(K-1)}t}\underline{V}^{(K-2)}(0)\cdots\underline{V}^{(0)} \\
&-\underline{U}^{(0)}(0)\cdots\underline{U}^{(K-2)}(0)\underline{V}^{(K-2)}(0)\cdots\underline{V}^{(0)}(0) \\
&+O(\epsilon) \qquad (2.6)
\end{aligned}
$$

where the final scale here (K-1) is such that $\underline{A}^{(K-1)}(0)$ has the same number of ergodic classes as $\underline{A}^{(k-1)}(\epsilon)$ for $\epsilon \in (0, \epsilon_o]$. Also $O(\epsilon)$ in (2.6) is uniform for $t \in [0, \infty)$. Thus we have a decomposition of our process in terms of increasingly aggregated processes describing behavior at longer and longer time scales. Note that the major contribution of [1] is the identification of the critical ϵ-dependent terms in $\underline{V}^{(k)}(\epsilon)$. In particular, higher order terms are of central importance for so-called almost transient states, i.e. states which at a particular time scale are transient for $\underline{A}^{(k)}(0)$, but are not transient for $\underline{A}^{(k)}(\epsilon)$ and therefore may provide critical paths between ergodic classes.

Consider now a simple FMS consisting of two machines, designated machines 1 and 2. Each of the machines is capable of operating on each of the two parts, type 1 and type 2. Machine 1 is flexible and unreliable. The flexibility indicates that the machine may operate on either part 1 or part 2 interchangeably without setting up. Therefore there is no set-up activity associated with this machine. It is, however, unreliable, indicating that it is subject to random failures and therefore there are failure and repair events defined for this machine, as well as a failure activity. Machine 2 is the opposite of machine 1, being reliable, but inflexible. To switch between part types, it is necessary to cease operations and perform the set-up

activity.

Essentially, by assuming that various events occur with exponential holding times, we can obtain a Markov chain model for this FMS. It is convenient to think of the state of this model as consisting of a set of components. These components are

- Failure status of machine 1 (failed, working)
- Set-up status for machine 2 (Set-up for part 1, switching to part 2, Set-up for part 2, switching to part 1)
- Machine 1 operations (working on part 1, working on part 2, idle)
- Machine 2 operations (working on part 1, working on part 2, idle)
- Decision Variable for Machine 1 (Loading decisions being made or not being made)
- Decision Variable for Machine 2 (Loading decision being made or not being made)

Obviously not all combinations of components make sense - - e.g. machine 2 cannot be working on part 2 if it is set up for part 1, and in fact this FMS has a 40-state model. Also, we include explicitly the notion of decision states to model loading decisions. That we associate exponential holding times with the time to make decisions is of no consequence (except to allow us to stay within the Markovian chain framework) as this will be the fastest process in the model which thus will be aggregated away at the first step of our procedure. The holding times and different rates associated with the decision components can be thought of as corresponding to scheduling decisions, and indeed, our long-term objective is to use the analysis presented here as the basis for designing control and scheduling schemes for FMS's.

The transitions of the multi-component state are also best thought of on an individual component basis, although some transitions change more than one component (e.g. a decision to begin an operation on a type 1 part changes the system to a non-decision state (since the decision has been completed) and changes the machine state from idle to operating on part 1) and the rates of changes for particular components depend on the other components (again we can't decide to begin operating on part 2 on machine 2 until it is set up for part 2). The complete set of elementary component rates for our model are

- Failure rate, P. This is the transition rate from machine 1 operating on part 1 to machine 1 failed (and of course idle).
- Repair rate, R. This is the transition rate from machine 1 failed to machine 1 working (but idle).
- Set up rate, S^{-1} This is the transition rate from machine 2 being set up for 1 or 2 to completely set up for that part type.
- Setup initiation rates, $F_s(i,j)$. These are the rates of transitions corresponding to setting machine 2 up for operation j when the failure status of machine 1 is i (i=0,1 correspond to failed and working respectively). Again, these should be thought of as scheduling parameters.
- Rates T_{ij}^{-1} at which operation j is completed on machine i.
- Decision completion rates L_{ij} for machine i resulting in the decision to initiate operation j. Again these are scheduling parameters and are in general functions of other components of the state. For example, if machine 2 is set up for part 1, we may use decision rates for machine 1 that favor operation 2.

The key features of an FMS on which our analysis rests is that the rates just described are of drastically different orders of magnitude. A reasonable ordering of the sizes of these rates is the following:

$$\begin{aligned} L_{ij} &= \lambda_{ij} \\ T_{ij}^{-1} &= \epsilon\tau_{ij}^{-1} \\ S^{-1} &= \epsilon^2 s^{-1} \\ F_s(i,j) &= \epsilon^2 f_s(i,j) \\ P &= \epsilon^3 p \\ R &= \epsilon^3 r \end{aligned} \tag{2.7}$$

where each of the lower case quantities are O(1). This equation implies that decisions characterize the fastest time scale, machine operations the next, then setups and set-up decisions, and finally at the slowest time scale, failures and repairs. With these choices, we now have a complete specification of a model of the form of (2.1). Applying the methodology described previously, we then obtain a sequence of models $\underline{A}^{(0)}(0)$, $\underline{A}^{(1)}(0)$, $\underline{A}^{(2)}(0)$, $\underline{A}^{(3)}(0)$ describing the dynamics of the FMS at different time scales:

$\underline{A}^{(0)}(0)$: at this time scale, the only transitions we see are decisions.

$\underline{A}^{(1)}(0)$: at this next time scale, decision transitions occur so quickly that their behavior can be averaged, yielding a model that captures the completion of operations over a scale at which neither set-up or failure events occur. Note that the averaging of the decision variable has the effect of reflecting the decisions to switch between the two parts on machine 1.

$\underline{A}^{(2)}(0)$: At this time scale individual part completions occur very frequently

and can be averaged. This scale focuses on set-up decisions and completions.

$\underline{A}^{(3)}(0)$: At this time scale, the focus is on failures and repairs. Again the averaging implied by the occurrence of faster events and captured by (2.5) leads to an effective failure rate reflecting the fact that in our model failures can only occur when machine 1 is operating on a part and not during times when it is idle.

III. Event Frequencies at Different Time Scales

In an FMS one is typically interested in production rates for different part types. In our simple example, production of a single part corresponds to the occurrence of one of several transitions in the Markov chain (e.g. machine 1 produces a type 1 part when its operational status changes from "operating on part 1" to "idle" -- the other components of the state, however, are not fixed, so there are several full state transitions corresponding to this completion event). It is also clear that the rate at which parts are produced depends on the time scale over which one looks at the process. For example, if one looks at production rates at a time scale commensurate with part production times, then these rates are the corresponding τ_{ij}^{-1} if machine j is operating on part i at the time (e.g. at this time scale, one sees no type 2 production if machine 1 is failed and machine 1 is set up for part 1). On the other hand, if one looks over a very long time period, in which there are many setups, failures and repairs, one could expect to see average production rates that take into account down time due to failures and setups as well as the scheduling parameters controlling part loading and set-up decisions.

What the preceding discussion suggests is another hierarchical approach in which rates at one time scale are averaged to produce rates at the next time scale. Note two interesting features of this concept: at faster time scales we are counting individual transitions in our models; at slower time scales, the individual transitions have been "aggregated away" and thus we are dealing with average numbers of transitions. Secondly, the development of a general theory for this type of computation is somewhat more delicate than the multiple time scale analysis described in the previous section. For example, suppose that there are two transitions, one from state i to state j and one from state m to state n, that correspond to the same physical event. Suppose further that the transition rate from i to j is ϵ, while the exact ergodic probability of state i is $\frac{1}{2}$; similarly, suppose that the rate from m to n is much larger, namely 1, but the ergodic probability of state m is ϵ. In this case, both of these possible transitions are of equal importance (the m to n transitions may occur more quickly, but we are in state m less frequently). State m in this example is an almost transient state so that if we looked at the ergodic probability matrices, $\underline{U}^{(k)}(0)$ introduced in the previous section we would find that they yielded a 0 probability of being in state m and hence would not account for the m to n transitions in computing the desired event rate.

To overcome this, let us first define ϵ-dependent versions of the probabilities in $\underline{U}^{(k)}(0)$. Specifically, let $\rho(t)$ denote the original Markov chain, let j be any state in the process corresponding to $\underline{A}^{(k)}(\epsilon)$ and let I denote any of the ergodic classes of $\underline{A}^{(k)}(0)$. Also, let Δt be a time interval that is explicitly a function of ϵ, i.e. so that

$$\epsilon^{-k} = o(\Delta t) \tag{3.1}$$

and

$$\Delta t = o(\epsilon^{-(k-1)}) \tag{3.2}$$

so that Δt is long with respect to the kth time scale, but short with respect to the (k+1)st. Then define

$$u_{jI}^{(k)}(\epsilon) = \lim_{\epsilon \to 0} \Pr\{\rho(t+\Delta t) \in j \mid \rho(t) \in I\}. \tag{3.3}$$

Note that $\{u_{jI}^{(k)}(0)\}$ are the elements of $\underline{U}^{(k)}(0)$. We then let $\tilde{u}_{jI}^{(k)}(\epsilon)$ denote the leading-order term of $u_{jI}^{(k)}(\epsilon)$, and define $\underline{\tilde{U}}^{(k)}(\epsilon)$ accordingly. In [2] an efficient method for calculating $\underline{\tilde{U}}^{(k)}(\epsilon)$ is described.

Suppose that we are interested in counting a certain set of transitions that all correspond to a common event. Specifically, for each state i in our process, let W_i denote the set of states j such that we wish to increase our count by 1 if an i to j transition occurs. Define the row vector $\underline{Q}^{(0)}(\epsilon)$ as

$$\underline{Q}^{(0)}(\epsilon) = \left[q_1^{(0)}(\epsilon), \cdots, q_N^{(0)}(\epsilon)\right] \tag{3.4}$$

(where $\underline{A}^{(0)}(\epsilon)$ is NxN) and

$$q_i^{(0)}(\epsilon) = \sum_{j \in W_i} a_{ji}^{(0)}(\epsilon). \tag{3.5}$$

Then define

$$\underline{Q}^{(k)}(\epsilon) = \underline{Q}^{(k-1)}(\epsilon)\underline{\tilde{U}}^{(k-1)}(\epsilon). \tag{3.6}$$

Let $\eta(t)$ denote the counting process corresponding to counting all of the transitions of interest. We then have the following

<u>Theorem:</u> Consider the time interval Δt satisfying

$$\epsilon^{-k+1} = o(\Delta t), \Delta t = o(\epsilon^{-k}), \tag{3.7}$$

then

$$\lim_{\epsilon \to 0} \frac{\left| q_I^{(k)}(\epsilon) - E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in I\right]\right|}{q_I^{(k)}(\epsilon)} = 0. \tag{3.8}$$

Furthermore, if

$$q_I^{(k)}(\epsilon) \geq O(\epsilon^{k-1}) \tag{3.9}$$

then the expectation in (3.8) can be dropped - - i.e. we have an almost sure quantity.

What this result states is that $q_I^{(k)}(\epsilon)$ is the leading-order term in the average count frequency at the kth time scale assuming we are in aggregate state I at that time scale. Furthermore, if enough transitions take place at this time scale - - i.e. if (3.9) holds, this asymptotic count frequency equals the observed count frequency almost surely.

The complete application of this result for the computation of effective production rates for our FMS example is described in [2]. We note here only the ultimate result. Specifically, if we look at Δt's longer that the slowest time scale - - i.e. a time period over which many failures and repairs occur - - we obtain a single effective production rate for each part type on each machine (at this time scale we have essentially aggregated our Markov chain into a single state). For example, the effective production rate for type 1 parts on machine 2 at this time scale has the form

$$u^{(4)} = \frac{p^{(3)}}{r + p^{(3)}} u^{(3)}(0) + \frac{r}{r + p^{(3)}} u^{(3)}(1) \tag{3.10}$$

where "(4)" denotes this very long time scale, while "(3)" denotes the prceding time scale. Here $p^{(3)}$ is the effective machine failure rate at this preceding time scale (recall failures can only occur when machine 1 is working on a part), while $u^{(3)}(0)$ and $u^{(3)}(1)$ are the total type 1 production rates when machine 1 is failed and working respectively (recall at this time scale the failure status is frozen). Thus (3.10) represents a weighting of these two rates by the percentages of time one expects machine 1 to be failed or working. If we back up another step we can relate $u^{(3)}(0)$ and $u^{(3)}(1)$ to a set of production rates at the second time scale at

which no setups are observed, and we will find that $u^{(3)}(0)$ and $u^{(3)}(1)$ are weighted combinations of these production rates with weights reflecting the amount of time machine 2 is set up for part 1. This process can be backed up further until we are back to the original τ_{ij}^{-1} rates.

References

[1] J.R. Rohlicek and A.S. Willsky. *The Reduction of Perturbed Markov Generators: An Algorithm Exposing the Role of Transient States.* Technical Report, MIT LIDS, 1986.

[2] C.A. Caromicoli. *Time Scale Decomposition Techniques for Flexible Manufacturing Systems.* Technical Report, MIT LIDS, 1987.