TIME SCALE ANALYSIS TECHNIQUES FOR FLEXIBLE MANUFACTURING SYSTEMS

by

CARL ADAM CAROMICOLI

B. Eng. Mgt., Electrical Engineering, McMaster University
(1986)

SUBMITTED TO THE DEPARTMENT OF
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1988

Signature of Author ______________________________
Department of Electrical Engineering and Computer Science
January 15, 1988

Certified by ______________________________
Alan S. Willsky
Thesis Supervisor

____
Stanley B. Gershwin
Thesis Supervisor

Accepted by ______________________________
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Time Scale Analysis Techniques For Flexible Manufacturing Systems

by

Carl Adam Caromicoli

Submitted to the Department of Electrical Engineering
and Computer Science on January 15,1988 in partial
fulfillment of the requirements for the degree of
Master of Science

Abstract

This thesis uses 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, a new theoretical result is presented on the computation of asymptotic rates of particular events in perturbed Markov processes, where an event may correspond to the occurence of one of several transitions in the process. This result is used to compute effective production rates at different time scales, taking into account the occurence of set-ups and failures.

Thesis Supervisors :

Alan S. Willsky
Title : Professor of Electrical Engineering

Stanley B. Gershwin
Title : Senior Research Scientist,
Laboratory for Manufacturing and Productivity

Acknowledgements

I would like to express my deep gratitude for the advice and help of my supervisors Alan S. Willsky and Stanley B. Gershwin over the last year and a half. Their help has been invaluable and I feel very fortunate to have worked with them for this time.

I wish to acknowledge the Air Force Office of Scientific Research for funding this research under Grant AFOSR-88-0032 and the Army Research Office under Grant DAAL03-86-K-0171. In addition I would like to thank the Natural Science and Engineering Research Council of Canada for their support.

I would also like to thank my parents for their help and understanding throughout the years. Without them, this opportunity at MIT would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Flexible Manufacturing Systems

A significant amount of interest has been generated regarding a production environment known as a Flexible Manufacturing System or FMS. Environments of this type hold promise for providing increases in productivity and a reduction in overhead in the manufacture of products. A distinguishing feature of Flexible Manufacturing Systems is that the machines or resources required for production are capable of performing many different operations, with a small amount of time needed to switch between operation types. This is in contrast to environments in which machines are dedicated to a single activity, sometimes involving only a single part type. The flexible environment offers opportunities for improvement in productivity because changes in product mixes, or the set of machines used to produce particular parts can be implemented without large expenditures of time or capital.

One of the consequences of the flexibility inherent in an FMS is the importance of the problem of scheduling and routing parts through the system. The large number of choices regarding which parts are to be machined, at what time, in what quantity and at which machine, makes the task of finding an optimal choice a very difficult and often numerically intractable analysis problem. In cases where several machines and part types are involved, even the analysis of production performance without control is infeasible. Therefore, in order to model, analyze and eventually control complex manufacturing systems, it is necessary to approximate the system in some fashion.

A number of approaches have been taken in modeling Flexible Manufacturing Systems. Most of these approaches fall into two categories. The first of these consists of combinatorial approaches in which one generates large models for even simple systems, but describes all of the significant events in a single model. Examples of this type of approach can be found in [2], [18], [22], and [25]. A second approach involves the construction of simpler models by describing only a small subset of important activities in the system. This assumption that certain activities in the system are relatively unimportant for analysis purposes is implicitly made by some authors in that only some events (such as machine failures) are introduced into the model. Examples of this type of approach are presented in [12] and [17]

Our approach is to use some plausible assumptions regarding the relative frequencies of events in the FMS to obtain a number of models which are valid at different time scales. Each of these models is far simpler than the model that we would obtain if we attempted to describe all events at once. For very large systems, a technique is described for combining this approach with a lumping technique which eliminates the details of certain dynamics in the system to further simplify the model.

In this work we use continuous time, finite state, Markov chains to model the system. Our motivation in using such a model for an FMS is threefold. First, we note that these models are already commonly used to describe events in analyses of Flexible Manufacturing Systems. For example, failure events are modeled using a Markov chain by both [20] and [12]. Secondly we note that there are already efficient methods available in the literature for performing time scale analyses on finite state Markov chains [23]. Finally, using this type of model, (Finite State Markov Chains) the key features of the approach can be clearly demonstrated without the extensive calculations required for more complex models (such as Semi-Markov models which do not require the assumption of exponentially distributed activity times). If such models are required to provide a better approximation to the real system, techniques exist in the literature to perform a decomposition [23]. The calculations require more effort, but the results will be similar in form to those which we obtain in this work.

1.2 Background

1.2.1 FMS Modeling and Analysis

A number of model types that have been used for FMS analysis are described by [26]. The methods described include techniques from queueing theory, perturbation analysis and mathematical programming. Queueing networks can be used for capacity analysis or the examination of interactions between system components which compete for limited resources. This information is useful in making planning decisions related to the set up of the FMS, such as the assignment of part types and operations to various machines. For complex systems this approach can become impractical and therefore simplifications to existing queueing models have been developed in [4] by Buzacott.

Mathematical programming techniques are useful for making scheduling decisions, which are those decisions related to the introduction of parts into a system and their subsequent routing. In [18] Lageweg et al attempt to solve a scheduling problem for a job shop using integer programming and bounding techniques; however, the approach cannot deal with reasonably sized systems. An integer programming approach to production planning in Flexible Manufacturing Systems was also developed by Stecke [25]. The method can handle problems of a realistic size, but works with a deterministic model only and a small number of activity types.

Perturbation analysis [13,14] is a means of quickly using a single simulation run to obtain results for configurations that differ from a basic system by some small amount. Sensitivities of system performance to changes in certain parameters can then be obtained. This technique is useful in making design decisions, which determine the number of machines and general layout in the FMS.

In [12], Hildebrant uses the queueing theory approach to analyze a system that experiences random failures which interrupt operations. The failures in the system are modeled by a continous time Markov Chain, and the time required to complete an operation is modeled as an exponentially distributed random variable. It is assumed that the events related to operations occur much more frequently than the failures, and therefore reach probabilistic equilibrium between changes in the failure state of the system. An algorithm for approximating the optimal routing policy and evaluating subsequent performance using Mean Value Analysis (MVA) is then presented. Unfortunately, the method suffers from the drawback of becoming quite complex for systems with several machines and several operations. In addition, there is no simple means of accounting for other activities in the FMS which may compete for the time of the resources or machines, such as setups required for changes between part families. The fact that the failures occur much less frequently than the operations themselves is used to justify the assumption of probabilistic equilibrium between failure events; however, a general means of reducing model complexity using this separation in frequency is not provided.

The time optimal control of a dynamic system with jump parameters is considered by [20]. This is an approach that could be classified as a mathematical programming approach as described by Stecke. Failures in the system are modeled by a continuous time Markov Chain. The goal of the analysis is to determine the optimal routing and loading decisions for the FMS. The method extends the techniques used for the control of systems with jump parameters by making a set of assumptions about the performance function and the solution itself. As stated in the paper, the method cannot be easily extended to handle general systems, and results in a computationally intractable problem, even for very simple systems.

As indicated above, many analysis techniques suffer from the problem of large increases in computational complexity when the systems being modeled are increased to a realistic size. Therefore many researchers have attempted to decrease the analytical complexity by generating hierarchical models which, for instance, explicitly separate the design, planning and scheduling decisions (or whatever classifications of decisions and activities they believe to be appropriate.)

Hutchinson [15] describes a hierarchical structure for the management and control of an FMS. The hierarchy is based upon the types of decisions that are made in the FMS and the lengths of duration for the events associated with these decisions. The three levels associated with the fastest events are very similar to those which will be generated by our analysis in Chapter 5. Specifically, these levels involve (in order of decreasing frequency) routing decisions, part loading decisions and decisions regarding the distribution of workload among the machines in the FMS. The first two levels involve routing and dispatching decisions, which correspond to the scheduling decisions described by Stecke, while the workload distribution issues are closely related to the planning decisions described by Stecke. The upper levels in the hierarchy are associated with management decisions regarding which parts will be produced and what resources will be added or deleted. Although the three lowest levels resemble the levels generated by our analysis, the paper does not provide a discussion of how to model the system or the decision processes themselves.

Buzacott [3] also discusses a hierarchical structure for modeling and analyzing Flexible Manufacturing Systems. His analysis utilizes a queueing system model, and incorporates a hierarchy with three levels. These levels are described as pre-release planning, input/release control, and operational control. These levels are similar to levels 3, 2 and 1 of the hierarchy described by Hutchinson. Buzacott discusses which events should have associated decision processes at the various levels. He does not however use the hierarchy to provide any means of simplifying the model or the associated analysis for large complex systems. The formulation of approximate, multi-level models is suggested as an important area of research.

The hierarchical structures developed in [17], [1], [11], [9] and [10], make explicit use of the types of events that may occur in a particular FMS and their differences in frequency to define levels in a hierarchy. In [17], a three level hierarchy is developed by Gershwin and Kimemia, which is somewhat similar in nature to the hierarchies of Hutchinson and Buzacott. The first level of the hierarchy deals with decisions over time intervals that are the same order of magnitude as the length of time required to complete an operation on a part, and is associated with what is called a sequence controller. The second level makes decisions over lengths of time which are the same order of magnitude as the time between machine failures. Decisions at this level are made by what are called the routing and flow controllers. Finally, the third level is associated with events affected by plans and policies of management, such as reconfigurations for different part families. In addition to describing the form of the hierarchy itself, [17] also provides rules that can be used to decide which actions controllers should take and when to take them, at the lower levels of the hierarchy. In [1], an actual problem is formulated and the performance analyzed, while [11] extends the work by including the effects of setups. The work of these papers is generalized in [9], which extends the hierarchical concepts to systems with an arbitrary number of different activities. The associated capacity constraints are discussed for a general k level hierarchy, as well as rules for making decisions.

The work in this thesis parallels the framework of [9,10], and uses similar terminology. This thesis deals only with the modeling issue, leaving control aspects to future work. The advantage of the approach taken by this work is that it provides a very general hierarchical structure by simply starting from a Markov Chain model and applying techniques from the theory of time scale decompositions of Markov chains. In addition, the resulting structured approach to determining the numerical measures of behavior provides a compact, precise set of relationships between the many variables that are present in the model. The analysis techniques that are presented are useful for the simplification and analysis of Markov Chain models. A brief overview of papers related to these topics in the literature is provided in the following subsection.

1.2.2 Markov Chain Analysis Techniques

In Chapter 4, a number of concepts are presented which are related to the simplification and analysis of Markov Chains. The techniques are related to three basic operations on a chain; aggregation, lumping and transition frequency calculation.

The concept of aggregating Markov Chains has been considered by several authors in the literature. It is basically a means of explicitly breaking the model up into models of lower complexity, each of which represents the system at a different time scale. Essentially, one can think of the chain as displaying a certain type of transition behavior if examined over short time intervals and a behavior that is quite different when examined over very long time intervals. An example of a system with behavior at multiple time scales can be taken from the manufacturing problem. Suppose that a machine completes an operation on a part once per minute while it is working, but fails every 2 to 3 days and once failed requires 1 day to be repaired. Examining the system over a 5 minute interval, we will see individual operations completed, but the machine will remain either failed or in working order with a very high probability. Examining the machine over a one week period however, will mean that individual operations are unimportant, but the failures and repairs will be observed.

Various approaches have been taken to the problem of aggregating Markov Chains. In [6], Courtois provides a method of aggregating systems described by

$$\underline{\dot{x}}(t) = \underline{A}(\epsilon)\underline{x}(t) \tag{1.1}$$

where $\underline{x}$ is a vector of probabilities for each state and $\underline{A}(\epsilon)$ is the transition rate matrix of the Markov Chain. This matrix contains elements which are a function of ϵ, the perturbation parameter which is a small positive number. He provides an intuitively pleasing and simple method of obtaining models for the chain at fast and slow time scales, but is not able to capture the correct behavior for certain classes of systems (those which contain sequences of rare transitions). Work done by Coderch et al [5] generalizes the decomposition to include all Markov chains, but at the expense of a much more complicated procedure. The work presented by Rohlicek in [23] obtains a compromise by developing an algorithm which captures the correct multiple time scale behavior for systems which cause Courtois' method to fail, but still retains an intuitively simple approach.

The relationship between aggregation and lumping of Markov Chains is discussed in [8] by Delebecque et al. Lumping is a method for reducing the size of a Markov Chain by combining the unneeded details that are present in the dynamics. The paper by Delebecque provides both a set of required conditions for a Markov Chain to be exactly lumpable and a procedure for calculating the transition rate matrix of the new, reduced dimension chain. Details regarding the lumping method are provided in Section 4.4.

Finally, the calculation of the expected frequency of state transitions has been examined under the name of probabilistic flow rates by Kielson in [16]. The approaches of Section 4.5 extend the basic concepts of expected transition frequencies to a multiple time scale model of a Markov chain. Furthermore, it is demonstrated that for the purposes of these calculations, the simplifications that are made in [23] may be too coarse to capture these expected frequencies. A modified approach, much in the spirit of Rohlicek and Willsky is developed to overcome this difficulty.

1.3 Structure of The Thesis

This thesis is comprised of 6 chapters, including Introduction and Conclusions. Chapters 2 and 3 deal with the basic problem of modeling a Flexible Manufacturing System as a continuous time Markov Chain. Chapter 2 starts out by providing a basis for the model that is generated in Chapter 3. The terminology and classification of entities in an FMS are consistent with the concepts introduced in [9,10]. Chapter 3 uses the framework discussed in Chapter 2 to obtain a Markov Chain model of a simple system that is introduced as an example. The example is deliberately chosen to be simpler than a model of a realistic system so that the important issues in the analysis are not obscured by the details of calculations.

Chapter 4 presents some techniques that can be used to analyze the type of model that is generated in Chapter 3. The three main concepts introduced in Chapter 4 are the aggregation and lumping of Markov Chains and the calculation of the expected frequencies of events that are modeled by the chain. The results for the first two techniques are taken from the literature with the details that are required for our purposes presented in Sections 4.3 and 4.4. The results for the event frequency calculations are new and are presented in Section 4.5, followed by derivations of the formulae.

The fifth chapter applies the techniques of Chapter 4 to the Markov Chain that was generated in Chapter 3. The numerical results obtained from a time scale decomposition, or aggregation of the model are compared to the qualitative expectations that one would obtain based upon the assumptions that were made in generating the model in Chapter 3.

Finally, the conclusions are presented in Chapter 6. This chapter also provides a list of suggestions for future research topics that show promise and which would enhance the usefulness of the concepts introduced by this thesis.

Chapter 2

The Manufacturing Environment

2.1 Introduction

The purpose of this chapter is to describe the types of manufacturing systems that we are modeling and which features of those systems we will try to characterize. The type of environment that we will be modeling is known as a Flexible Manufacturing System or FMS. The terms and concepts regarding the features and events in such a system follow those introduced in [9,10] by Gershwin.

2.2 Flexible Manufacturing

The term flexible manufacturing refers to an environment in which a group of machines are available to perform various operations on parts that the system produces. Each of the machines is capable of performing a number of different operations on one or more of the parts. The term flexible refers to the fact that a machine can switch between operations within a given family of operations with negligible or no set-up time. In general, the operations that are performed by these machines can be classified as discrete tasks, each with a clearly defined start and finish time.

For systems in which there are a large number of machines operating on a large number of parts, the problem of analyzing and optimally controling the system becomes very difficult. The models of the FMS become so complex that a numerical analysis becomes intractable for systems of realistic size. Therefore there has been an emphasis on developing simple models which accurately account for many of the important events in the systems. One possible means of reducing the complexity of a model is to ignore some of the events or activities in the FMS which are deemed to be the least important. This approach may be appropriate for some manufacturing systems, but the events which are unimportant for one FMS may be important for another. A second approach is to assume that the activities that take place in the system are independent. For example, we might assume that the failure of one machine does not affect the activities at another machine. Again, this approach may be appropriate for some environments, but is unsatisfactory when the actions that are carried out in the FMS vary significantly as a function of what has taken place in the past. The second situation is much more likely to resemble the behavior in a realistic system. For example, if one machine fails in an FMS, one might expect there to be an increase in production on other machines to compensate for the loss of the first machine.

This work forms a large model that incorporates the resources, activities and events described by Gershwin. The first model that is formed describes all of the dynamics of the system with a single Markov Chain and therefore suffers from the problem of complexity described above. The means of reducing this complexity involves a decomposition of the complete model using hierarchical concepts. A series of models is generated, each corresponding to a particular level in the hierarchy.

The result is that the individual models describe only a subset of all of the activities that take place, but the dependence of activities on one another is preserved. The consideration of a small subset of the activities associated with each level is possible because the model for each subset is associated with a different time scale. Therefore, at a given level in the hierarchy, events which occur much less frequently may be regarded as static. Conversely, the fine details of the more frequent events are blurred away, so that only an aggregate view of these events is required.

2.3 Modeling Framework

As stated in the previous subsection, the model employed in this work parallels the hierarchical framework for Flexible Manufacturing Systems that was established by Gershwin. This section presents some of the concepts provided in his work which are relevant to the model that we develop.

Within the FMS there are machines available to perform the operations required for part production. The machines are part of a broader class of items known as *resources.* These resources are necessary for the production process, but are not consumed by it. The modeling techniques presented in this work are generalizable to any resources; however, our demonstrative example deals only with simple machines, which are capable of performing a small number of operations.

The second set of features of importance are defined by Gershwin to be *activities.* Due to the discrete nature of the system we are modeling, each activity has a clearly defined start and finish time. In addition, each activity is associated with a resource or in our case a machine. For example, machine A may start an activity such as cutting a piece of metal at time X and finish at time Y. The act of cutting the metal is known as an activity at machine A, while the time required to complete the activity, Y-X, is known as the activity's *duration*. It is important to note that only some of the activities in the system are voluntary. For example, operations on parts are voluntary because we can decide whether or not a machine will initiate a given operation. The failures of a machine, however, are not voluntary because the machines are unreliable and therefore we cannot eliminate failures or predict exactly when they will occur.

We may also define a quantity known as the *frequency* of an activity. The frequency is defined simply to be the total number of occurances of an activity on a time interval, divided by the length of the time interval. Therefore, if activity B is performed 24 times by resource (or machine) A between 8 a.m. and 4 p.m., the frequency of activity B on resource A is 3 per hour. Gershwin defines the variable u_{ij} to be the frequency of type j activities on machine i. The notation defined in Chapter 5 is consistent with this definition.

Finally, the *occupation* of a machine or resource may be defined. The occupation of a machine by an activity is the fraction of time that the resource is engaged the activity.

2.4 Activities in a Flexible Manufacturing System.

Having defined the general concepts of resources and activities in Section 2.2, we now describe the specific cases of these features that are encountered in this work. In the case of resources this is simple because we model only a single resource, the machines themselves. For general models; however, other resources might be required such as transportation vehicles, storage locations or operators. In each case, the object or person is required for production, but is not consumed in the process.

The activities that may be modeled are somewhat more complex. Among the more important activities are machine failures, machine setups, operations, and decisions. Machine failures start with a failure event and finish with a repair. The failure is any uncontrollable event that makes the machine incapable of performing in a normal fashion. The times when these events occur are not under the control of the operators, while the time required to repair the machine may or may not be under the operators' control.

The set-up activity is required because a machine can be configured to perform only a finite set of operations at a given time. This is because each operation may require a different tool, and there is limited on-line storage provided by the tool magazine. Therefore, if a decision is made at some time to perform an operation requiring tools from a different set, the set-up activity must be initiated. While the machine is being prepared for a different set of operations, it is incapable of producing parts. Once the machine has been set up for the new operation, the set-up activity is complete.

An operation on a part is also classified as an activity. The start of the operation occurs when a part is loaded and ends when the part is removed. The decision process is defined as the activity that starts when an operation is complete and ends when a new operation begins. This activity has a very short duration relative to the time required to complete an operation because its length is just the time required to make a decision regarding which part to load next. Note that we define an operational mode for the time when the machine is sitting idle, but capable of production.

Additional activities may be modeled in a similar manner, namely, by defining an event associated with the start of an activity and an event associated with the completion. For example, if we are to model maintenance activities, the starting event would be when an individual first commences maintenance activities on a machine or resource, disabling its normal operation. The completion event will then be the point at which normal operations are resumed.

2.5 Relative Frequencies of Activities or Events

The previous section defined the frequency of an activity as the number of occurrences of the activity per unit time. Therefore we may associate a frequency with each activity that occurs in the system. The basic premise on which the hierarchical framework rests is that the activities which occur in an FMS will have frequencies which differ substantially in magnitude. This assumption is described by equation (7) in Gershwin [1987]. Basically he groups the activities in the system into sets J_1, J_2, ..., J_k such that the activities in each set have frequencies f_1, f_2, ..., f_k of different magnitude. The assumption defined in Gershwin is given by (2.1).

$$f_1 << f_2 << ... << f_k \tag{2.1}$$

Therefore, any pair of events from different groups are assumed to occur at significantly different rates. This concept underlies the hierarchical framework, because the hierarchy assumes that between two occurrences of an activity, all of the more frequent events reach steady state, while the less frequent events do not occur at all. Therefore, at any particular level in the hierarchy, the details of the dynamics in the system that occur at other levels are irrelevant.

2.6 Summary

In this chapter, a basic discussion of Flexible Manufacturing Systems and their characteristics has been presented. Emphasis was placed on the framework and definitions introduced by Gershwin as they apply to this work. In additon to the terminology, the assumptions upon which his hierarchical framework is based are repeated, as they provide a basis for the analysis which is performed in Chapter 5. The chapter provides a partial list of activities that may occur in an FMS, including all activities which are introduced in the models of this work. Finally, a discussion of the hierarchical framework suggested by the assumptions regarding activity frequencies is presented. In Chapter 5, we see how our analysis of a basic model, starting with the frequency assumption, naturally generates this framework.

Chapter 3

Markov Chain Model of an FMS

3.1 Introduction

This chapter casts the description of an FMS in Chapter 2 into the framework to be analyzed here. In order to clearly demonstrate the results, we develop a model for a simple Flexible Manufacturing System. The characteristics of this system are such that the important features of the analysis procedure are demonstrated. In order to generate a model for a realistic system, a greater number and variety of events is required, but using such an example would only obscure the most important results.

3.2 Characteristics of the System

The system that we will be modeling is represented diagramatically in Figure 3.1 ([10]). The resources of the system consist 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, in Figure 3.1. Machine 1 is flexible and unreliable. The flexibility indicates that the machine may operate on either part 1 or part 2 interchangeably

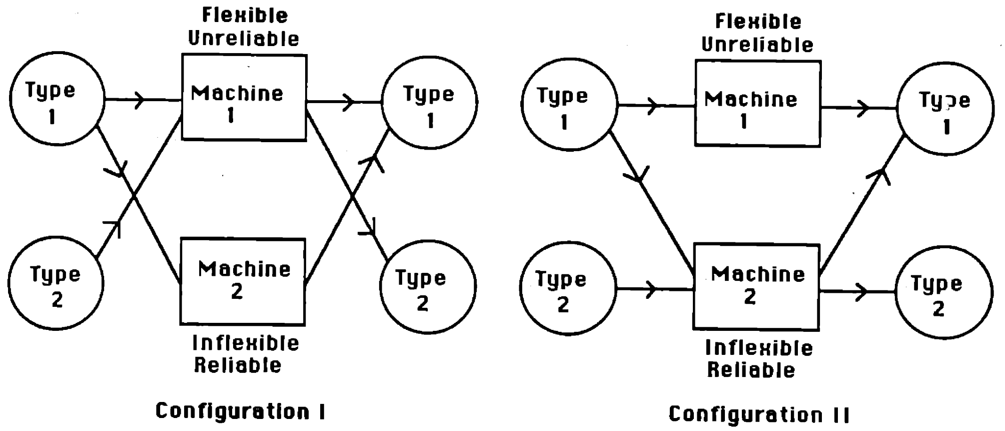


Figure 3.1: Simple FMS

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. The inflexibility is depicted in Figure 3.1, which shows that this machine may operate only on part type 1 or only part type 2. To switch between parts (configurations I and II), it is necessary to cease operations and perform the set-up activity.

For machine 1 we model the failure events as occurring at random, uncontrollable times, with the time required to complete the repair also random. Both of these times are modeled as possessing exponential distributions. The time until a setup is initiated is also modeled as an exponential random variable. A few comments are required here regarding the exponential distribution for the time until setups are initiated. The exponential distribution is an idealization of the true behavior for this and in fact all of the events in the model. It is used because it greatly simplifies the computational aspects of the time scale analysis. However, from the work of Rohlicek [23] involving semi-Markov processes, the form of the distribution that we select does not affect the final result that is obtained for the time scale decomposition. Therefore, instead of obscuring the intuitive aspects of our results with more involved calculations, we use an exponential distibution for such quantities as the time until a setup is initiated, rather than modeling them as random variables with alternate distributions or even as deterministic quantities. The time required to complete the setup is also modeled as an exponential random variable.

The operation and decision activities are described in Chapter 2. An operation starts when a part is loaded and ends when the part is removed. In addition we define the operational status of a machine that is not operating on a part as idle. The decision activity takes place between the operations. The time between decisions (either the time required to complete an operation or the time spent idle) is also modeled as an exponential random variable.

3.3 State Space Model

The model that will be employed to describe the FMS is a continuous time, finite state Markov chain. Our first step is to define a state space for the chain. In order to obtain a state space, we start by defining the concept of a mode and a component. We define a *mode* as the condition of a resource, which is changed by a particular type of event. For example, if we have a machine that experiences failures and repairs, we may define two failure modes for the machine, one for when the machine is failed and one for the machine in working order. A *component* of the state is comprised of the set of modes associated with a particular characteristic of the resource. For our machine which experiences failures and repairs, one of the the components of the state is the failure component, which may be in one of the two

Component	Description	Symbol	Modes
1	Failure of Machine 1	α	0,1
2	Setup of Machine 2	σ	1-4
3	Operation by Machine 1	γ_1	0-2
4	Operation by Machine 2	γ_2	0-2
5	Decision for Machine 1	β_1	0,1
6	Decision for Machine 2	β_2	0,1

Table 3.1: Modes For Simple FMS

modes described above. The six components which define the state of the system that we are modeling are listed in Table 3.1.

In order to simplify our description of the combinations of activities, we define a set of modes for each activity in Table 3.1. Starting with activity 1 there are two modes determined by whether machine 1 is failed or in working order. We assign $\alpha = 0$ for the case when the machine is failed and $\alpha = 1$ for the working order case. The set-up component for machine 2 is associated with 4 modes. The machine may be prepared to operate on parts 1 or 2 ($\sigma = 3$ or 4) or may be preparing to operate on parts 1 or 2 ($\sigma = 1$ or 2).

The operational components (3 and 4 in Table 3.1) are each associated with 3 modes, 1 for each of part 1 and 2 ($\gamma_i = 1$ and 2 respectively) and an additional mode when the machine is idle ($\gamma_i = 0$). The fifth and sixth components in Table 3.1, the decision components, each have two possible modes, with 1 mode for when a decision is being made ($\beta_i = 1$) and another for all other times ($\beta_i = 0$).

The state of the system could be defined as the Cartesian product of the six components. Given the number of modes that are possible for each component, this scheme would generate 288 unique modal combinations for our simple model. We note, however, that only a fraction of these 288 combinations are necessary to model all of the conditions that may be experienced by the real system. Table 3.2 defines

Modeling Assumptions	Restriction on Modal Combinations
1 Machine 1 cannot operate unless it is in working order	If $\alpha = 0$ then γ_1 and β_1 must be 0.
2 Machine 2 is inflexible	If $\sigma = 1$ then γ_2 cannot be 2. If $\sigma = 2$ then γ_2 cannot be 1.
3 Machine 2 cannot operate while being set up	If $\sigma = 1$ or 2 then γ_2 and β_2 must be 0.
4 Machines are idle while decisions are made.	If $\beta_i = 1$ then γ_i must equal 0.

Table 3.2: Conditions for Elimination of Meaningless States

which combinations are necessary to describe the system, by listing the conditions which are required for a particular combination of modes to be meaningful. Each entry in the table lists a modeling assumption which is based on our attempt to accurately model the FMS while maintaining simplicity. The table also lists the restrictions that the assumptions place on the valid modal combinations.

For example, the third assumption requires that if the system is in set-up modes 1 or 2 (corresponding to setting up machine 2 for parts 1 or 2 respectively) then machine 2 cannot operate on a part. Therefore a set-up mode of 1 or 2 ($\sigma = 1$ or 2) implies operational and decision modes of 0 ($\beta_2 = \gamma_2 = 0$), which describes the system when there are no operations being performed or decisions being made.

Combining the restrictions listed in Table 3.2, we find that 40 states are necessary to model our system. These states and their components are listed in Table 3.3.

State	α	σ	γ_1	γ_2	β_1	β_2	State	α	σ	γ_1	γ_2	β_1	β_2
1	0	1	0	0	0	0	21	1	3	0	1	1	0
2	0	2	0	0	0	0	22	1	3	0	1	0	0
3	0	3	0	0	0	1	23	1	3	1	0	0	1
4	0	3	0	0	0	0	24	1	3	1	0	0	0
5	0	3	0	1	0	0	25	1	3	1	1	0	0
6	0	4	0	0	0	1	26	1	3	2	0	0	1
7	0	4	0	0	0	0	27	1	3	2	0	0	0
8	0	4	0	2	0	0	28	1	3	2	1	0	0
9	1	1	0	0	1	0	29	1	4	0	0	1	1
10	1	1	0	0	0	0	30	1	4	0	0	1	0
11	1	1	1	0	0	0	31	1	4	0	0	0	1
12	1	1	2	0	0	0	32	1	4	0	0	0	0
13	1	2	0	0	1	0	33	1	4	0	2	1	0
14	1	2	0	0	0	0	34	1	4	0	2	0	0
15	1	2	1	0	0	0	35	1	4	1	0	0	1
16	1	2	2	0	0	0	36	1	4	1	0	0	0
17	1	3	0	0	0	0	37	1	4	1	2	0	0
18	1	3	0	0	1	0	38	1	4	2	0	0	1
19	1	3	0	0	0	1	39	1	4	2	0	0	0
20	1	3	0	0	0	0	40	1	4	2	2	0	0

Table 3.3: Listing of the State Space

3.4 State Transitions

Now that we have a state space for the model, we need to define the dynamics. The model that we employ is a finite state, continuous time Markov chain. To describe the dynamics we first need to determine which transitions are possible. We note that each transition event causes more than one of the components of the state to change. This is because the components of the state are not totally independent as we saw in the discussion of Table 3.2. Therefore, an event must correspond to a change in a number of the components of the state, so that a new meaningful combination of modes is reached following the transition.

For example, consider the event corresponding to a failure of machine 1. Initially, when machine 1 is in working order, it is possible to have operational modes 0,1 or 2, but if it is in the failed condition, we may only have operational mode 0, because the machine cannot perform operations on parts. Therefore the failure event causes both a change the failure component from mode 1 to mode 0, and a reset to 0 of the operational mode.

Table 3.4 summarizes the effect of each possible event on the components of the state. Consider row 3 of Table 3.4. This row indicates that the initiation of a setup causes the set-up component (σ) to change to 1 if it was 4 or to 2 if it was 3 before the setup. In addition, the operation and decision modes are reset to 0 regardless of the starting mode, because operations cannot be completed while the machine is being set up. Similarly, consider row 5 of the table. We see that before a new operational mode is initiated, the operational component is idle. After the initiation of a new operational mode, the machine may be operating on parts 1 or 2 (operational modes 1 and 2), or not operate at all (idle mode 0). In addition, the decision component changes from mode 1 (making a decision) to mode 0 (not

Event	α	σ	γ_1	γ_2	β_1	β_2
1 M1 Fails	$1 \rightarrow 0$	-	$1,2 \rightarrow 0$	-	-	-
2 M1 Rep'd	$0 \rightarrow 1$	-	-	-	$0 \rightarrow 1$	-
3 Setup Start	-	$4 \rightarrow 1$ or $3 \rightarrow 2$	-	rst 0	-	rst 0
4 Setup End	-	$2 \rightarrow 4$ or $1 \rightarrow 3$	-	-	-	$0 \rightarrow 1$
5 M1 Start	-	-	$0 \rightarrow 0-2$	-	$1 \rightarrow 0$	-
6 M1 End	-	-	rst 0	-	$0 \rightarrow 1$	-
7 M2 Start	-	-	-	$0 \rightarrow 0-2$	-	$1 \rightarrow 0$
8 M2 End	-	-	-	rst 0	-	$0 \rightarrow 1$

Comments The event column lists each of the eight possible events that can occur. *M1 Fails* represents the event when machine 1 fails while *M1 Rep'd* is the event that occurs when the machine is repaired. *Setup start* corresponds to the initiation of the set-up activity on machine 2, the end of which is marked by a *Setup end.* *M1 Start* and *M2 Start* correspond to the start of a new operational mode for a machine while *M1 End* and *M2 End* indicate the end of an operational mode. An entry $A \rightarrow B$ in the table indicates that the indicated component changes from mode A to mode B for the corresponding event. A *rst* 0 entry indicates that regardless of the initial mode of the component, the mode following the event will be 0. Finally, a dash indicates that the event has no effect on the mode of that component.

Table 3.4: Changes in Components Due to Various Events

making a decision). Therefore the machine is idle during the decision process, but once a decision is made regarding the next operational mode, the machine either remains idle by choice or starts operating on one of the parts.

Table 3.4 and the associated discussion determine which transitions can occur, but we must also define the rate at which they occur. For a Markov chain, all transition times are exponentially distributed. If the rate for a transition in a Markov chain is R, then the time until a transition occurs has a distribution given by (3.1),

$$p(t) \;=\; R\, e^{-Rt}, \tag{3.1}$$

and the mean time until a transition is given by $\frac{1}{R}$.

We now assign symbols to the physical quantities associated with each of the events in the system. Starting with the failure events, the rate at which failures occur will be called P. Equivalently, the mean operational time until a machine fails is P^{-1}. In a similar fashion, we define a repair rate R and the rates at which setups are initiated, $F_s()$. The parentheses indicate that arguments may be required to differentiate among a number of different setup initiation rates. For the setups themselves we let S be the mean time required to set up a machine. Therefore the setup completion rate is S^{-1}.

We note that these symbols may represent the transition rate for a number of transitions in the chain. For example, the repair rate, R, is the transition rate for any transition between two states that results from a repair. In the case of setups, there may be several rates at which setups are initiated, depending on the current failure and set-up modes. Therefore, we must define a number of setup initiation rates.

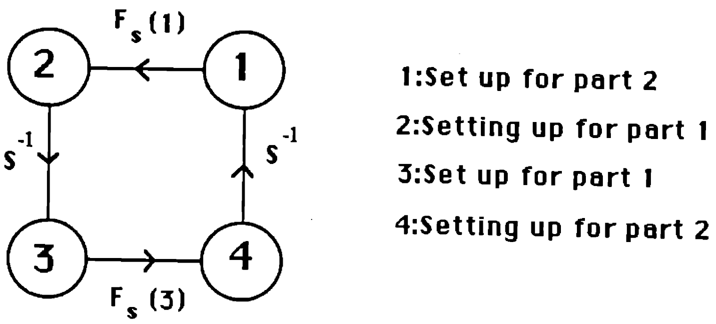


Figure 3.2: Example of set-up dynamics

Consider the Markov Chain shown in Figure 3.2. In this chain, we assume that a transition from state 1 to state 2 corresponds to the initiation of a setup for one part class while transitions from state 3 to state 4 correspond to the initiation of a setup for a second part class. The transitions from state 2 to 3 and 4 to 1 correspond to the completion of these setups. The two set-up initiation rates may be different and therefore are labelled $F_s(1)$ and $F_s(3)$, where the arguments indicate the state of origin for the transition. When we form the transition matrix for our main example in Section 3.5, we use the the failure and set-up modes of the state of origin to differentiate between set-up rates.

The mean time to completion of operation j on machine i is T_{ij} and therefore the operation completion rate is ${T_{ij}}^{-1}$. The decision completion rates are L_{ij} for machine i if the resulting decision is to initiate operation j. We may also define several operation initiation rates, using the failure and set-up modes of the state of origin of the transition to distinguish between transitions, just as we did for

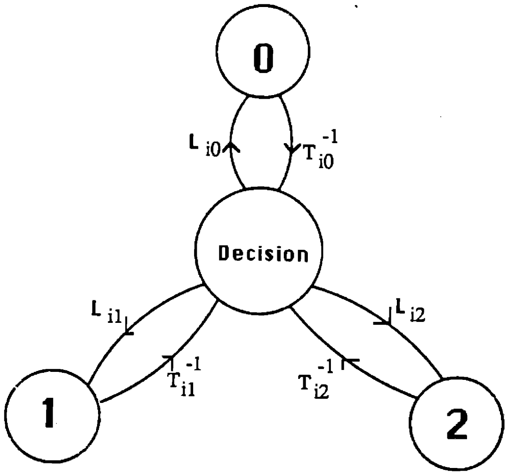


Figure 3.3: Operation and Decision Dynamics for Machine i

setups. The need to differentiate among a number of operation initiation rates arises because the choice operations for a machine may vary for different failure and set-up modes of the two machines. Therefore, a set of operation initiation rates is defined for each unique combination of set-up and failure modes.

The operation and decision dynamics are shown in Figure 3.3. The center state in the figure corresponds to the state of the system when the decision regarding which operation to complete next is being made, while the other 3 states correspond to operational modes 0,1 and 2. Operational mode 0 as described previously, corresponds to the situation when the machine sits idle. In order to maintain consistency with the other parts of the model and to simplify calculations, we model the time spent in the idle operational mode until another operational decision is initiated as an exponential random variable, with mean value T_{i0}. As in the case of setups, the exponential distribution has no effect on the time scale decomposition results, but by reducing the amount of computation, provides a simpler exposition of the qualitative aspects of the analysis.

State	Failure Component	Operational Component
1	Failed	Idle
2	Working	Idle
3	Working	Busy

Table 3.5: State Listing by Component

3.5 Formation of the Transition Matrix

Given that we are modeling an FMS as a continuous time Markov Chain, we can describe the dynamics using equation (3.2),

$$\dot{\underline{x}}(t) = \underline{A}\ \underline{x}(t) \tag{3.2}$$

where $\underline{\mathrm{x}}$(t) is the vector of state probabilities and $\underline{\mathrm{A}}$ is the rate transition matrix for the chain. Since the system contains 40 states, $\underline{\mathrm{A}}$ is a matrix of dimension 40 by 40. We therefore choose to work with an even simpler system initially (3 states) to more clearly demonstrate the principles involved in forming the transition matrix. Having completed this we will proceed to form the 40 by 40 matrix for our example.

Three State Example

Suppose we have a chain with 3 states having components defined by Table 3.5. In accordance with our previous discussion, the failure mode is changed by failure and repair events, and the operational component is changed by an operation completion (with the idle condition defined as an operation). Using the symbols P for the failure rate, R for the repair rate and T_o, T_1 for the mean operation completion times, we can draw the state transition diagram shown in Figure 3.4.

Failures in this model, as in our 40 state model, only occur when the machine is operating on a part, and therefore result in a transition from state 3 to state

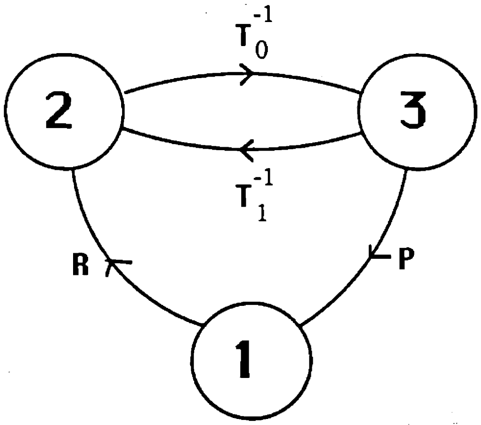


Figure 3.4: State Transition Diagram for Three State Example

1. The repairs cause a transition from state 1 to state 2 because the system is always returned to idle following a repair. Finally, operation completions result in transitions between states 2 and 3, corresponding to the idle and busy operational modes.

Given the diagram of Figure 3.4, we can form the transition rate matrix:

$$\underline{A} = \begin{bmatrix} * & 0 & P \\ R & * & {T_1}^{-1} \\ 0 & {T_0}^{-1} & * \end{bmatrix}. \tag{3.3}$$

This matrix is very easy to form directly due to the small number of states. This will not be the case for models that are even as simple as our 40 state example. Therefore we must develop a method of partitioning the state space and the transition rate matrix so that it can be described in portions. This is accomplished by selecting a subset of the components of the state and assigning states to individual subspaces which contain unique combinations of modes for those components. For example, consider our three state chain in Figure 3.4. If we select the failure component of the state as a basis for our partitioning, we realize that there are two possible modes and hence we will have two subspaces. This first subspace, corresponding to the failed mode is designated F, where F={1}, and the second subspace, representing machine 1 in working order is W, where W={2,3} (refer to Table 3.5). The transition matrix for the partitioned system can be written in the form

$$\underline{A} = \begin{bmatrix} \underline{A}_{FF} & \underline{A}_{FW} \\ \underline{A}_{WF} & \underline{A}_{WW} \end{bmatrix} \tag{3.4}$$

where:

$$\underline{A}_{FF} = [-R] \tag{3.5}$$

$$\underline{A}_{WF} = \begin{bmatrix} R \\ 0 \end{bmatrix} \text{(Repair Transitions)} \tag{3.6}$$

$$\underline{A}_{FW} = [0\ P]\ \text{(Setup Transitions)} \tag{3.7}$$

$$\underline{A}_{WW} = \begin{bmatrix} -T_0{}^{-1} & T_1{}^{-1} \\ T_0{}^{-1} & -T_1{}^{-1} - P \end{bmatrix} \text{(Operational Transitions)} \tag{3.8}$$

Each submatrix of $\underline{A}$ contains transition rates representing a single event type.

Formation of 40 State Transition Matrix

We proceed to form the transition matrix for our 40 state model by immediately partitioning the state space. Selecting subspaces which have common failure and set-up modes, we can use Table 3.3 to obtain eight subspaces for our model. The states belonging to each of the subspaces are listed in Table 3.6.

The components of the state that should be chosen as a basis for our partitioning (for example the failure component was selected for the three state chain), should be those which change least frequently. This results in transitions between subspaces

Subspace	Member States
1	1
2	2
3	3,4,5
4	6,7,8
5	9,10,11,12
6	13,14,15,16
7	17 to 28
8	29 to 40

Table 3.6: Partitioned State Space

which are small in magnitude compared to the transitions within the subspaces. Based on assumptions that will be made in Section 3.6, this will result in very small numbers of transitions between subspaces, i.e. they will occur at a very slow rate, and hence the transitions between the subspaces can be handled separately. For our model, the failure and setup related events are assumed to be the least and second least frequent events. Therefore, the failure and set-up components are chosen as a basis for the partitioning. The assumptions regarding the magnitudes of the frequencies are described in more depth in Section 3.6.

The matrix $\underline{A}$ can then be written in the partitioned form described by (3.9).

$$\underline{A} = \begin{bmatrix} \underline{A}_{11} & \cdots & \underline{A}_{18} \\ \vdots & \ddots & \vdots \\ \underline{A}_{81} & \cdots & \underline{A}_{88} \end{bmatrix} \tag{3.9}$$

We can now proceed to define the elements of each submatrix of $\underline{A}$. The rates for the decision and setup initiation events are denoted by $L_{ij}(\alpha, \sigma)$ and $F_s(\alpha, \sigma)$ to indicate explicit dependence on the failure and set-up modes.

The diagonal submatrices are given by

$$\underline{A}_{11}\ ,\ \underline{A}_{22} = [0], \tag{3.10}$$

$$\underline{A}_{33} = \begin{bmatrix} * & {T_{20}}^{-1} & {T_{21}}^{-1} \\ L_{20}(0,1) & * & 0 \\ L_{21}(0,1) & 0 & * \end{bmatrix}, \tag{3.11}$$

$$\underline{A}_{44} = \begin{bmatrix} * & {T_{20}}^{-1} & {T_{22}}^{-1} \\ L_{20}(0,2) & * & 0 \\ L_{22}(0,2) & 0 & * \end{bmatrix}, \tag{3.12}$$

$$\underline{A}_{55} = \begin{bmatrix} * & {T_{10}}^{-1} & {T_{11}}^{-1} & {T_{12}}^{-1} \\ L_{10}(1,3) & * & 0 & 0 \\ L_{11}(1,3) & 0 & * & 0 \\ L_{12}(1,3) & 0 & 0 & * \end{bmatrix}, \tag{3.13}$$

$$\underline{A}_{66} = \begin{bmatrix} * & {T_{10}}^{-1} & {T_{11}}^{-1} & {T_{12}}^{-1} \\ L_{10}(1,4) & * & 0 & 0 \\ L_{11}(1,4) & 0 & * & 0 \\ L_{12}(1,4) & 0 & 0 & * \end{bmatrix}, \tag{3.14}$$

$$\underline{A}_{77} = \begin{bmatrix} * & {T_{20}}^{-1} & {T_{10}}^{-1} & 0 & {T_{21}}^{-1} & 0 & {T_{11}}^{-1} & 0 & 0 & {T_{12}}^{-1} & 0 & 0 \\ L_{20} & * & 0 & {T_{10}}^{-1} & 0 & 0 & 0 & {T_{11}}^{-1} & 0 & 0 & {T_{12}}^{-1} & 0 \\ L_{10} & 0 & * & {T_{20}}^{-1} & 0 & {T_{21}}^{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & L_{10} & L_{20} & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ L_{21} & 0 & 0 & 0 & * & {T_{10}}^{-1} & 0 & 0 & {T_{11}}^{-1} & 0 & 0 & {T_{12}}^{-1} \\ 0 & 0 & L_{21} & 0 & L_{10} & * & 0 & 0 & 0 & 0 & 0 & 0 \\ L_{11} & 0 & 0 & 0 & 0 & 0 & * & {T_{20}}^{-1} & {T_{21}}^{-1} & 0 & 0 & 0 \\ 0 & L_{11} & 0 & 0 & 0 & 0 & L_{20} & * & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & L_{11} & 0 & L_{21} & 0 & * & 0 & 0 & 0 \\ L_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & {T_{20}}^{-1} & {T_{21}}^{-1} \\ 0 & L_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & L_{20} & * & 0 \\ 0 & 0 & 0 & 0 & L_{12} & 0 & 0 & 0 & 0 & L_{21} & 0 & * \end{bmatrix}, \tag{3.15}$$

where the symbol L_{ij} is used as a short form for $L_{ij}(1,1)$. Finally

$$\underline{A}_{88} = \begin{bmatrix} * & T_{20}^{-1} & T_{10}^{-1} & 0 & T_{22}^{-1} & 0 & T_{11}^{-1} & 0 & 0 & T_{12}^{-1} & 0 & 0 \\ L_{20} & * & 0 & T_{10}^{-1} & 0 & 0 & 0 & T_{11}^{-1} & 0 & 0 & T_{12}^{-1} & 0 \\ L_{10} & 0 & * & T_{20}^{-1} & 0 & T_{22}^{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & L_{10} & L_{20} & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ L_{22} & 0 & 0 & 0 & * & T_{10}^{-1} & 0 & 0 & T_{11}^{-1} & 0 & 0 & T_{12}^{-1} \\ 0 & 0 & L_{22} & 0 & L_{10} & * & 0 & 0 & 0 & 0 & 0 & 0 \\ L_{11} & 0 & 0 & 0 & 0 & 0 & * & T_{20}^{-1} & T_{22}^{-1} & 0 & 0 & 0 \\ 0 & L_{11} & 0 & 0 & 0 & 0 & L_{20} & * & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & L_{11} & 0 & L_{22} & 0 & * & 0 & 0 & 0 \\ L_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & T_{20}^{-1} & T_{22}^{-1} \\ 0 & L_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & L_{20} & * & 0 \\ 0 & 0 & 0 & 0 & L_{12} & 0 & 0 & 0 & 0 & L_{22} & 0 & * \end{bmatrix} \tag{3.16}$$

where the symbols L_{ij} represent $L_{ij}(1,2)$ in this case.

Note that these matrices contain only elements which represent the rates for the decision and operation completion events. This is because the events related to the set-up and failure activities generate transitions between the subspaces of Table 3.6, because each subspace represents constant failure and repair modes. Therefore, the elements of the off- diagonal submatrices correspond to these rates. In particular the submatrices given by (3.17) and (3.18) contain elements which are either zero or equal to the failure rate.

$$\underline{A}_{15} \text{ , } \underline{A}_{26} = [0\ 0\ P\ P] \tag{3.17}$$

$$\underline{A}_{37}, \underline{A}_{48} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & P & 0 & 0 & P & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & P & 0 & 0 & P & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & P & 0 & 0 & P \end{bmatrix} \tag{3.18}$$

Consider the definition of $\underline{A}_{15}$ in equation (3.17). This matrix describes the transition rates from states in the fifth subspace to the states of the first subspace. The first subspace contains state 1 only, while the fifth subspace contains states 9 through 12 from Table 3.6. Equation (3.17) indicates that the transition rates from states 9 and 10 to state 1 are zero because failures cannot occur when the system is in these states (machine 1 is not operating). Conversely the transition rates from states 11 or 12 to state 1 are P. The repair rates in equations (3.19) and (3.20) are the transition rates from states for which machine 1 is failed to states for which machine 1 is in working order.

$$\underline{A}_{51} = \underline{A}_{62} = \begin{bmatrix} R \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{3.19}$$

$$\underline{A}_{73} = \underline{A}_{84} = \begin{bmatrix} R & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & R & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & R & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T \tag{3.20}$$

Equations (3.21) to (3.25) define the transition rates between states that are in subspaces with different set-up modes. The rates at which setups are initated while machine 1 is failed are given by

$$\underline{A}_{14} = [F_s(0,2)\ F_s(0,2)\ F_s(0,2)] \tag{3.21}$$

and

$$\underline{A}_{23} = [F_s(0,1)\ F_s(0,1)\ F_s(0,1)], \tag{3.22}$$

while the corresponding rates when machine 1 is in working order are

$$\underline{A}_{58} \ , \ A_{67} \ = \begin{bmatrix} F & F & 0 & 0 & F & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & F & F & 0 & F & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & F & F & F & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & F & F & F \end{bmatrix}, \qquad (3.23)$$

where $F \ = \ F_s(1,2)$ for $\underline{A}_{58}$ and $F_s(1,1)$ for $\underline{A}_{67}$. The reason for the increased size of the matrix in (3.23) is that there are additional operational modes generated in the case where machine 1 is working. The setup completion rates are given by

$$\underline{A}_{31} \ , \ \underline{A}_{42} \ = \begin{bmatrix} S^{-1} \\ 0 \\ 0 \end{bmatrix} \qquad (3.24)$$

when machine 1 is failed and by

$$\underline{A}_{75} \ , \ \underline{A}_{86} \ = \begin{bmatrix} S^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & S^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & S^{-1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & S^{-1} & 0 & 0 \end{bmatrix}^{T} \qquad (3.25)$$

when it is in working order. All of the submatrices that are not defined by (3.10) through (3.25) are identically zero. The elements in these matrices represent rates for one step transitions that cannot occur because they would correspond to the simultaneous occurence of a failure event and a set-up event, which is modeled as being impossible.

3.6 Magnitudes of the Transition Rates

In Chapter 2 we stated an assumption, described by (2.1) that the magnitudes of the frequencies of events in an FMS are significantly different from each other. In addition, Subsection 3.5 assumes that failures and setup events have the smallest rates. The first assumption is necessary for our analysis in order to generate behavior at multiple time scales, but the second is not. Any ordering of the magnitudes of the frequencies which happens to provide a good approximation to the system we are modeling is sufficient for our techniques to be applied. For the purposes of illustration, we pick the ordering of magnitudes given by (3.26).

$$P, R << S^{-1}, F_s(i,j) << T_{ij}{}^{-1} << L(i,j) \tag{3.26}$$

This ordering agrees with the case A, defined by equation (22) in Gershwin [10]. The physical situation that this assumption corresponds to could be described as follows. The decisions in the system may take only a very short time to complete, say, a fraction of a second. The operations take the next shortest time to complete, on the order of 1-2 minutes. The mean time to initiate or complete a setup is longer, 1-2 hours, and finally the mean operation time until a failure (as well as the time to repair the machine) is 2-3 days.

In order to make the ordering of the magnitudes explicit in the model, we introduce the parameter ϵ, which is a small positive number which we call the perturbation parameter. Recognizing that for small ϵ,

$$1 >> \epsilon >> \epsilon^2 >> \cdots \tag{3.27}$$

we make the assumption that:

$$\begin{aligned} L_{ij}(m,p) &= O(1) \\ {T_{ij}}^{-1} &= O(\epsilon) \\ S^{-1}, F_s(i,j) &= O(\epsilon^2) \\ P, R &= O(\epsilon^3). \end{aligned} \tag{3.28}$$

Furthermore, we introduce a set of lower case quantities which are related to the variables in (3.28) by (3.29).

$$\begin{aligned} L_{ij}(m,p) &= \lambda_{ij}(m,p) \\ {T_{ij}}^{-1} &= \epsilon {\tau_{ij}}^{-1} \\ S^{-1} &= \epsilon^2 s^{-1} \\ F_s(i,j) &= \epsilon^2 f_s(i,j) \\ P &= \epsilon^3 p \\ R &= \epsilon^3 r \end{aligned} \tag{3.29}$$

Each of the lower case quantities are O(1). The expressions on the right side of (3.29) will be used in place of the upper case variables in the transition matrix. For example, (3.17) becomes

$$\underline{A}_{15} = [0\ 0\ \epsilon^3 p\ \epsilon^3 p], \tag{3.30}$$

where dependence of the elements of the matrix on ϵ is explicitly indicated.

3.7 Summary

In this chapter, we introduced a simple manufacturing system that displays some of the characteristics described in Chapter 2. A set of assumptions regarding the timing of events is made to enable us to model the system as a continuous time Markov Chain. The model that is developed is described in terms of a state and a set of transition rates. The transition rates are obtained from fundamental quantities describing the system that we are modeling (upper case).

The assumption of wide frequency separations is introduced explicitly into the model by using a parameter ϵ. A new set of variables is introduced (lower case), each of which are related to the physical rates by an integer power of ϵ. Finally, the rates corresponding to physical quantities are replaced in the transition matrix by expressions containing powers of ϵ and the $O(1)$ quantities defined in equation (3.29).

Chapter 4

Analysis Techniques for Markov Chain Models

4.1 Introduction

In Chapter 3 a Markov chain model of a simple flexible manufacturing system was formulated. The reason for employing this type of model is that various techniques can be used to reduce the dimensionality of the model and extract useful information about its behavior. The purpose of this chapter is to describe analytical tools for both the simplification of the model and the extraction of important information. These tools are applied to the FMS in Chapter 5.

We start with a set of examples in Section 4.2 which provide an introduction to the kinds of calculations that will be performed. Sections 4.3 and 4.4 describe two techniques that have been developed elsewhere which are useful for reducing the dimensionality of the model: aggregation and lumping. In the case of lumping, the number of states is reduced by eliminating aspects of the model which have no effect on the overall results we hope to obtain from our analysis. Aggregation

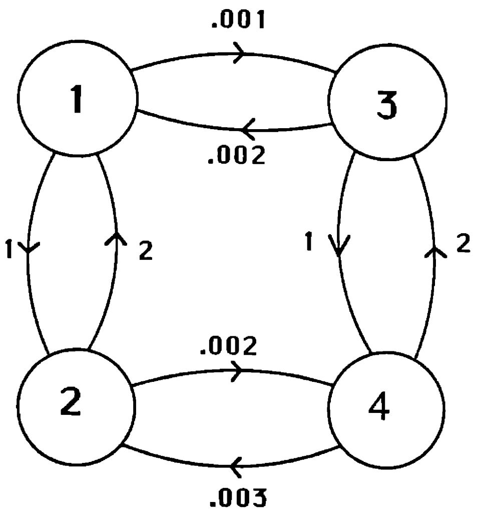


Figure 4.1: Motivating Example 1

also achieves a reduction in dimension; however, it is achieved by replacing a higher order model by several lower order models, without eliminating significant features of the dynamics.

Section 4.5 deals with the extraction of information from the model. Specfically, formulae for determining the expected frequency of events in the system being modeled are developed and presented. Finally, the appendix to the chapter provides a means of working with systems which have a large state space. The technique uses the multiple time scale behavior of the system to reduce the dimension of the matrices which are manipulated.

4.2 Motivating Examples

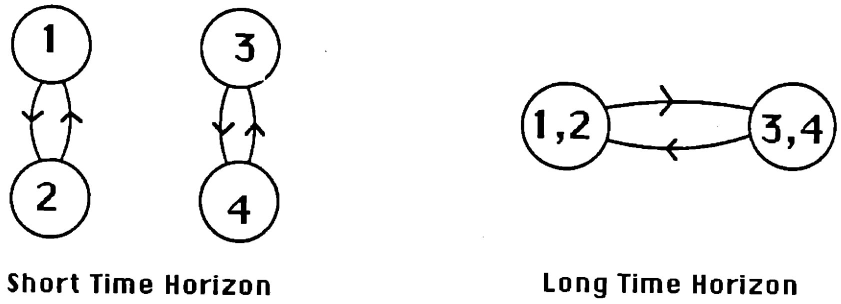
In this section we discuss the simplification and analysis techniques of interest as they apply to the simple system shown in Figure 4.1. The first simplification technique of interest is known as *aggregation*. This approach to simplification of the chain is also known as a *time scale decomposition*, because it relies on the fact that the system being studied exhibits different behavior if observed at different time scales.

Consider the continuous time Markov chain that is depicted in Figure 4.1. The transitions between the top and bottom states occur much more frequently than those from left to right or right to left. If we examine the system over a 5 to 10 unit time interval, we will very likely see several top/bottom transitions, but there is a very low probability that any left/right transitions will occur. If we examine the system over a time interval of 5000 units, a very large number of top/bottom transitions will occur; so many that individual transitions will become blurred. The number of left/right transitions however is likely to be on the order of 10 to 15, and therefore individual transitions can be distinguished.

The time scale decomposition technique relies on the fact that the total behavior can be approximated as a combination of these two short and long term dynamics, which are described by models of reduced dimension. The approximate model therefore exhibits dynamics which are a combination of the dynamics of the chains in Figure 4.2.

Unlike the aggregation or time scale decomposition approach to model simplification, *lumping* techniques do not approximate all of the dynamics of the original model. That is, some of the detail of the original model is lost and therefore lumping may be used only when part of the dynamics are not required in our analysis.

For example, suppose we are interested only in whether the system of Figure 4.1 is in the top pair of states (1,3) or the bottom pair of states (2,4). (Note that the top/bottom transition rates are independent of whether the system is in the left



Short Time Horizon

Long Time Horizon

Figure 4.2: Time Scale Decomposition Models

or right pair of states.) The lumping procedure could be used in this situation to combine states 1 and 3 and states 2 and 4 to obtain the model in Figure 4.3. The information regarding whether the system is on the left or right has been lost, but we have assumed that this information is not required for our analysis.

The information that is obtained by analyzing the model may take various forms. The analysis technique of Section 4.5 deals with the calculation of the frequency at which certain events take place in the system. For example, suppose we are interested in how frequently top to bottom transitions occur in the chain depicted by Figure 4.4. Recalling the time scale approximation concepts discussed above, over short time intervals we will remain on the left or the right for the duration of the interval with high probability. Therefore one of the two chains shown in Figure 4.5 will be an appropriate model.

Intuitively, we suspect that the frequency of top to bottom transitions will be different for the left and right cases, although exact numerical verification is left until Section 4.5. If we examine the system over very long time peroiods, we no longer remain exclusively in the left or right pair of states. For sufficiently long

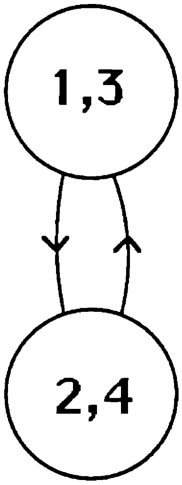


Figure 4.3: Lumped version of Motivating Example 1

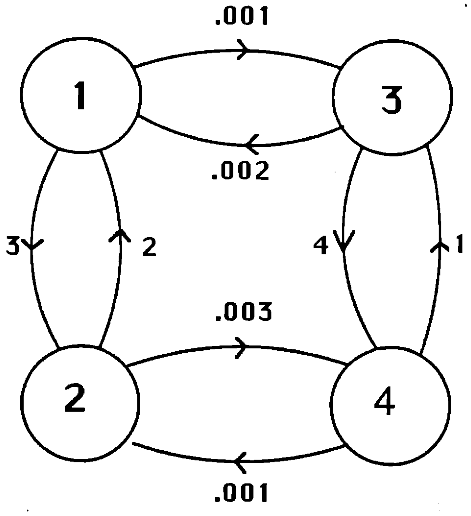


Figure 4.4: Motivating Example 2

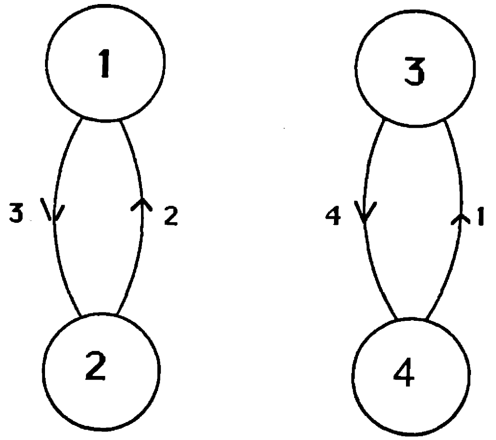


Figure 4.5: Example 2 over Short Time Intervals

intervals of time the system will spend some fraction of the interval on the left and the rest on the right. Therefore we expect that over this long interval of time, there will not be two separate top-to-bottom transition frequencies, but instead a single average frequency which is a function of the amount of time spent on each side. We proceed now with more detailed descriptions of the analytical tools described above.

4.3 Aggregation / Time Scale Decomposition

4.3.1 Specific Example

As outlined in Chapter 1, the problem of aggregating Markov Chains has been handled by several authors in the literature. The approach that is described here and which is used in subsequent calculations is developed in [23]. The technique can be applied to systems with multiple time scale behavior, such as the system of Figure 4.4, which is redrawn in Figure 4.6, with the quantity ϵ which we refer

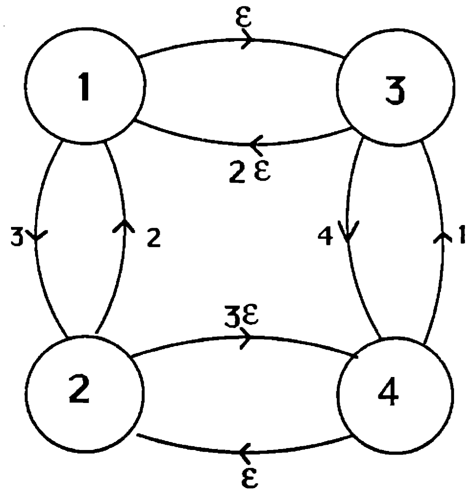


Figure 4.6: Example 2 Using Perturbation Parameter

to as the *perturbation parameter*. (To make the systems of Figures 4.4 and 4.6 equivalent, set $\epsilon = 0.001$.) In general, the techniques that follow require that ϵ be a small positive number.

The first step in the decomposition procedure is to obtain a system of the form illustrated in Figure 4.2 for short time scales. In our previous discussion we obtained that model by setting to zero the probability that a slow transition occurs. This effect can be achieved by setting $\epsilon = 0$ in Figure 4.6 so that left/right transitions do not occur. The resulting short time scale model is shown in Figure 4.7.

To obtain the model for long time scales, we must use the fact that many top/bottom transitions will occur between each left/right transition. Therefore, we assume that the top/bottom dynamics reach a steady state between each left/right transition and hence the aggregate left/right transition rates will be weighted averages of the rates while in the top and bottom states. The fraction of time spent in the top or bottom state can be approximated using the following ergodic probabilities for the short-time horizon model of Figure 4.7:

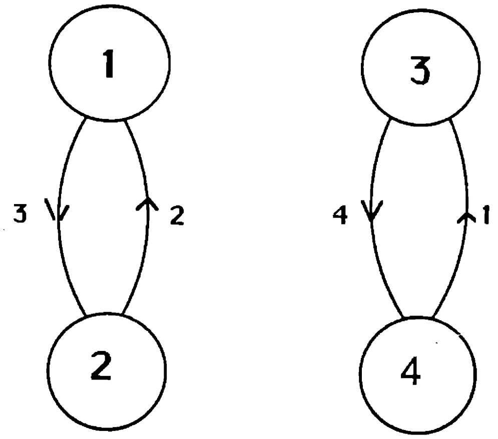


Figure 4.7: Example 2 Over Short Time Interval.

$$Pr\{State\ 1 \mid Left\} = \frac{2}{5} \tag{4.1}$$

$$Pr\{State\ 2 \mid Left\} = \frac{3}{5} \tag{4.2}$$

$$Pr\{State\ 3 \mid Right\} = \frac{1}{5} \tag{4.3}$$

$$Pr\{State\ 4 \mid Right\} = \frac{4}{5}. \tag{4.4}$$

Using these probabilites, we can calculate the left to right transition rates which are appropriate for this time scale. Specifically, if μ_1' is the aggregate left to right transition rate, then:

$$\begin{aligned} \mu_1' &= \epsilon Pr\{State\ 1 \mid Left\} + \ 3\epsilon Pr\{State\ 2 \mid Left\} \\ &= \frac{11}{5}\epsilon \end{aligned}$$

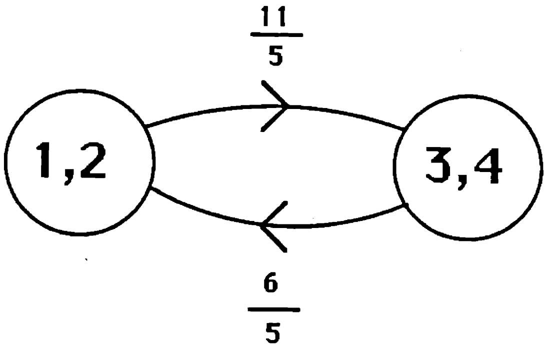


Figure 4.8: State Transition Diagram at Slow Time Scale

Similarly, if μ_2' is the right to left transition rate then:

$$\mu_2' = \frac{6}{5}\epsilon \tag{4.5}$$

Finally, we can make the change in time scale explicit by defining a new time variable τ given by:

$$\tau = \epsilon t \tag{4.6}$$

where t is the original time variable. The left/right transition rates at this new time scale become:

$$\mu_1 = \frac{11}{5} \tag{4.7}$$

$$\mu_2 = \frac{6}{5} \tag{4.8}$$

Using these results, we draw the state transition diagram for the chain at the slow time scale in Figure 4.8.

The results of Rohlicek [23] show that the combination of the slow dynamics of Figure 4.8 and the fast dynamics of Figure 4.7 constitute an asymptotically exact description of the original dynamics as $\epsilon \to 0$. In addition, the paper provides an algorithmic approach to the above calculations for a general chain. This approach is described in the following section.

4.3.2 Formalized Aggregation Approach

The paper by Rohlicek [23] provides a five step procedure for finding the aggregated version of a Markov chain and hence a model of the system at various time scales. The algorithm is set up so that it can be used recursively, each time using the most recently obtained, aggregated model, as the new starting point. In order to demonstrate the approach, we use the Markov chain introduced in Section 4.3.1, and shown in Figure 4.6. The probabilistic evolution of this process is given by:

$$\dot{\underline{x}}(t) = \underline{A}^{(0)}(\epsilon)\ \underline{x}(t) \tag{4.9}$$

where $\underline{\mathrm{x}}$(t) is the a vector of the probabilities of being in each of the four states at time t and $\underline{A}^{(0)}(\epsilon)$ is the transition rate matrix. For this example, $\underline{A}^{(0)}(\epsilon)$ is given by:

$$\underline{A}^{(0)}(\epsilon) = \begin{bmatrix} * & 2 & 2\epsilon & 0 \\ 3 & * & 0 & \epsilon \\ \epsilon & 0 & * & 1 \\ 0 & 3\epsilon & 4 & * \end{bmatrix} \tag{4.10}$$

with the superscript 0 indicating that we are at the initial time scale and the argument, ϵ, indicating dependence of the matrix on the perturbation parameter. The elements on the diagonals, abbreviated by asterisks, are such that the sum of the elements in each column is zero.

The first step of the algorithm is to separate the states into the ergodic and transient classes that are generated when ϵ is equal to zero. This will result in a transition matrix that possesses a block diagonal form if there are no transient states. For example, consider the system that was shown in Figure 4.7. Equation (4.10) becomes:

$$\underline{A}^{(0)}(0) = \begin{bmatrix} * & 2 & 0 & 0 \\ 3 & * & 0 & 0 \\ 0 & 0 & * & 1 \\ 0 & 0 & 4 & * \end{bmatrix} \tag{4.11}$$

If there were transient classes, then (4.11) would have had additional columns of non-zero elements and rows of zero elements corresponding to the transient states.

The second step of Rohlicek's algorithm requires the calculation of ergodic probabilities for each state, conditioned on the system being in each ergodic class. These probabilities are calculated with $\epsilon = 0$. Note that the ergodic probability of being in a state, conditioned on being in an ergodic class to which it does not belong, is zero. For example, in the chain shown in Figure 4.7,

$$Pr\{State\ 1 \mid Right\} = 0 \tag{4.12}$$

These conditional probabilities are then used to form the ergodic probability matrix, $\underline{U}^{(0)}(0)$, given by:

$$\left[\underline{U}^{(0)}(0)\right]_{iJ} = u_{iJ}{}^{(0)}(0) \tag{4.13}$$

where $u_{iJ}{}^{(0)}(0)$ is the ergodic probability of being in state i, given that the system is in ergodic class J. The ergodic probabilities of transient or almost transient states are 0 for $\epsilon = 0$.

We can now form the ergodic probability matrix for our example:

$$\underline{U}^{(0)}(0) = \begin{bmatrix} \frac{2}{5} & 0 \\ \frac{3}{5} & 0 \\ 0 & \frac{1}{5} \\ 0 & \frac{4}{5} \end{bmatrix} \qquad (4.14)$$

The third step of the algorithm is to form what is known as the class membership matrix, denoted $\underline{V}^{(0)}(\epsilon)$. The argument ϵ indicates that some of the elements of the matrix may be functions of ϵ. The elements of this matrix are $v_{J_i}{}^{(0)}(\epsilon)$, where $v_{J_i}{}^{(0)}(\epsilon)$ is the probability of J being the first ergodic class that the system enters, given that the system starts in state i. For a recurrent state (at $\epsilon = 0$), the probability is 1 for the ergodic class to which it belongs and zero otherwise. In the case of transient states, transitions may occur into more than one ergodic class, in which case there is more than one non-zero entry corresponding to that state. In all instances, the terms which are lowest order in ϵ are kept, while higher order terms may be discarded [Rohlicek, 1986].

For our example, there are no transient states and two ergodic classes, with states 1 and 2 in the first ergodic class and states 3 and 4 in the second ergodic class. The class membership matrix is therefore given by (4.15).

$$\underline{V}^{(0)}(\epsilon) = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \qquad (4.15)$$

The fourth and final step for the aggregation is to calculate the rate transition matrix for the model at the next slower time scale. From Rohlicek, this matrix is given by (4.16),

$$\underline{A}^{(1)}(\epsilon) = \frac{1}{\epsilon}\, \underline{V}^{(0)}(\epsilon)\, \underline{A}^{(0)}(\epsilon)\, \underline{U}^{(0)}(0), \qquad (4.16)$$

which for our example yields

$$\underline{A}^{(1)}(\epsilon) = \begin{bmatrix} \frac{-11}{5} & \frac{6}{5} \\ \frac{11}{5} & \frac{-6}{5} \end{bmatrix}. \tag{4.17}$$

In general, the procedure can be repeated, using $\underline{A}^{(1)}(\epsilon)$ as the new starting point; however, in our case, $\underline{A}^{(1)}(\epsilon)$ does not contain any $O(\epsilon)$ terms and therefore will not exhibit any additional dynamics over longer time scales.

Once the aggregation procedure has been repeated enough times to obtain models for each time scale at which the system displays dynamics, the original full dimensional chain can be approximated by a combination of the low order models. The analytical approximation to the solution of (4.9) takes the form of equation (4.18).

$$\begin{aligned} e^{\underline{A}^{(0)}(\epsilon)t} = \; & e^{\underline{A}^{(0)}(0)t} + \\ & \underline{U}^{(0)} e^{\underline{A}^{(1)}\epsilon t} \underline{V}^{(0)}(\epsilon) - \underline{U}^{(0)}\underline{V}^{(0)}(\epsilon) + \\ & \underline{U}^{(0)}\underline{U}^{(1)} e^{\underline{A}^{(2)}\epsilon^2 t} \underline{V}^{(1)}(\epsilon)\underline{V}^{(0)}(\epsilon) - \underline{U}^{(0)}U^{(1)}\underline{V}^{(1)}(\epsilon)\underline{V}^{(0)}(\epsilon) + \\ & \vdots \\ & \underline{U}^{(0)}...\underline{U}^{(k-2)} e^{\underline{A}^{(k-1)}\epsilon^{(k-1)}t} \underline{V}^{(k-2)}(\epsilon)...\underline{V}^{(0)}(\epsilon) - \underline{U}^{(0)}...\underline{U}^{(k-2)}\underline{V}^{(k-2)}(\epsilon)...\underline{V}^{(0)}(\epsilon) + \\ & O(\epsilon) \end{aligned} \tag{4.18}$$

For our example this becomes:

$$\underline{x}(t) = e^{\underline{A}(\epsilon)t}\,\underline{x}(0) \tag{4.19}$$

and

$$
\begin{aligned}
exp\{\underline{A}(\epsilon)t\} &= exp\left\{\begin{matrix} -3 & 2 & 0 & 0 \\ 3 & -2 & 0 & 0 \\ 0 & 0 & -4 & 1 \\ 0 & 0 & 4 & -1 \end{matrix}\right\} t + \\
&\begin{bmatrix} \frac{2}{5} & 0 \\ \frac{3}{5} & 0 \\ 0 & \frac{1}{5} \\ 0 & \frac{4}{5} \end{bmatrix} exp\left\{\begin{bmatrix} \frac{-11}{5} & \frac{6}{5} \\ \frac{11}{5} & \frac{6}{5} \end{bmatrix} \epsilon\, t\right\} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} - \\
&\begin{bmatrix} \frac{2}{5} & \frac{2}{5} & 0 & 0 \\ \frac{3}{5} & \frac{3}{5} & 0 & 0 \\ 0 & 0 & \frac{1}{5} & \frac{1}{5} \\ 0 & 0 & \frac{4}{5} & \frac{4}{5} \end{bmatrix} + O(\epsilon) \qquad (4.20)
\end{aligned}
$$

4.4 Lumping Techniques for Markov Chains

4.4.1 Specific Example

In Section 4.3, we saw that we could reduce a high order model to an approximation of that model composed of several lower order models which describe the system at different time scales. In our example, described by (4.10), the original model consists of a 4-state system, while the model obtained through aggregation techniques contains three 2-state systems. For chains with a larger number of states, the reduction in dimensionality of the largest chains that must be analyzed is significant.

The aggregation procedure may not, however, sufficiently reduce the size of the state space for models of complex systems with many types of dynamics. If this is

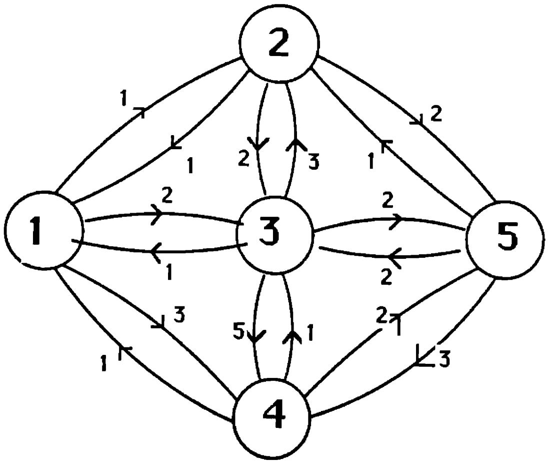


Figure 4.9: Example for Lumping Technique

the case, a procedure known as *lumping* may be used to further reduce the size of the state space. For example, suppose we have the Markov Chain shown in Figure 4.9. Comparing the rates leaving states 2,3 and 4 and entering state 1, we see that each of the rates are equal, with a similar condition holding for transitions entering state 5. Regardless of which state (2,3 or 4) the system is in, the transition rates out of the state are identical. In addition, we recognize that the total transition rates into the lumped state are the sums of the individual rates. Therefore, if we compare the Markov Chain depicted in Figure 4.9 with the chain in Figure 4.10, we see that the dynamics external to the subspace containing states 2,3 and 4 are identical.

Therefore, if we assume that for the purposes of our analysis the details of the dynamics within that subspace are unimportant, the chain shown in Figure 4.10 is an *exact* model of reduced dimension for the original system.

4.4.2 Formalized Lumping Approach

A structured approach to the lumping of Markov Chains and the conditions under which such an approach is possible are provided by Delebecque et al [8]. The

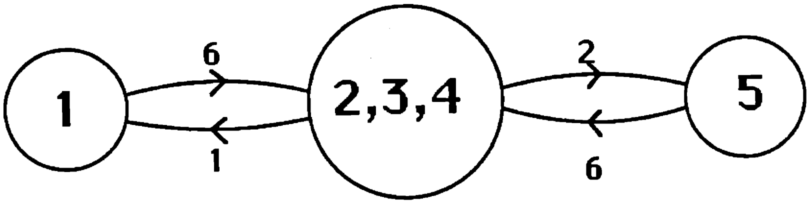


Figure 4.10: Lumped Chain

technique starts with a chain described by equation (4.21),

$$\underline{\dot{x}}(t) = \underline{A}\ \underline{x}(t) \tag{4.21}$$

where $\underline{A}(\epsilon)$ is an n by n matrix and $\underline{\mathrm{x}}$(t) is an n by 1 column vector if there are n states in the chain. The vector $\underline{\mathrm{x}}$(t) is a vector of the probabilities of being in each of the states. The goal of the lumping procedure is to find a new Markov Chain with fewer states than the original chain, and whose vector of state probabilities is $\underline{\mathrm{y}}$(t). The states of the lumped chain should be combinations of the states of the original chain, i.e. we should have

$$\underline{y}(t) \ = \ \underline{C}\ \underline{x}(t), \tag{4.22}$$

where $\underline{C}$ is called the lumping matrix. The matrix $\underline{C}$ is similar to the class membership matrix used in the time scale decomposition in that the elements c_{ij} must be of the form:

$$c_{ij} = \begin{cases} 1 & \text{if state j of the original chain is lumped} \\ & \text{into state i of the new chain} \\ 0 & \text{otherwise} \end{cases}$$

We define the transition matrix for the new chain to be $\underline{A}_c$, so that:

$$\dot{\underline{y}}(t) = \underline{A}_c \, \underline{y}(t). \tag{4.23}$$

The condition for lumpability that is obtained by Delebecque is that the matrices $\underline{C}$ and $\underline{A}_C$ exist such that

$$\underline{A}_c \, \underline{C} = \underline{C} \, \underline{A}. \tag{4.24}$$

Delebecque also provides a means of calculating the matrix $\underline{B}$, using

$$\underline{B} = \underline{W}\underline{C}(\underline{C}\underline{W}\underline{C}')^{-1} \tag{4.25}$$

and

$$\underline{W} = diag(w_1, ..., w_n). \tag{4.26}$$

If the chain is lumpable, then the values $w_1, ..., w_n$ can have any non-negative values such that at least one of the weights associated with each lumped state is greater than 0. Once we have defined $\underline{B}$, we can find $\underline{A}_c$ from (4.27).

$$\underline{A}_c = \underline{C} \, \underline{A} \, \underline{B} \tag{4.27}$$

Generally, the values w_i reflect a weighting of the states. The weighting is based on their relative contribution to the transition rate out of the lumped state. Therefore, a typical set of weightings would be their ergodic probabilities. In the lumpable case, the weighting does not affect the result for $\underline{A}_c$. For our simple example of Section 4.4.1, we can write the transition matrix as:

$$\underline{A} = \begin{bmatrix} * & 1 & 1 & 1 & 0 \\ 1 & * & 3 & 0 & 1 \\ 2 & 2 & * & 1 & 2 \\ 3 & 0 & 5 & * & 3 \\ 0 & 2 & 2 & 2 & * \end{bmatrix}, \tag{4.28}$$

and since we chose to lump states 2,3 and 4 we have,

$$\underline{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{4.29}$$

Now if we choose $\underline{W}$ as

$$\underline{W} = diag\{1,1,0,0,1\} \tag{4.30}$$

then

$$\underline{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.31}$$

and

$$\underline{A}_c = \begin{bmatrix} * & 1 & 0 \\ 6 & * & 6 \\ 0 & 2 & * \end{bmatrix} \tag{4.32}$$

The state transition diagram corresponding to equation (4.30) is identical to that provided in Figure 4.10.

4.5 Event Frequency Calculations

Sections 4.3 and 4.4 presented two techniques that can be used to reduce the dimension of a Markov Chain. Given that we have a reduced-dimension model, we must decide on what information is to be obtained from it and how it is to be calculated.

One type of information that may be of interest is the frequency at which particular events take place within the system. These frequencies are the number of events per unit time. For our Markov Chain model, the events in the system are represented by state transitions. The following subsections provide a compact method for calculating the expected frequency of these transitions. Subsection 4.5.1 provides a motivating example, 4.5.2 demonstrates a structured approach to the calculations, 4.5.3 demonstrates the approach in algorithmic form, 4.5.4 justifies the approach in 4.5.3 and 4.5.5 presents some results regarding the ergodicity of the Markov chain model and the effect on the expected frequency of transitions over very long time intervals.

4.5.1 Motivating Example

We start with the example that was illustrated in Figure 4.6. In this example, assume that we are interested in the frequency of the event which is represented by a transition from state 1 to state 2. From the discussion in Section 4.3, the chain of Figure 4.6 will exhibit different behavior at different time scales. Therefore, we expect that the transition frequencies may also be defined differently, depending on the time scale of observation. We start by considering time intervals Δt which are o(ϵ^0) = o(1). This notation indicates that

$$\lim_{\epsilon\to 0} \frac{\Delta t}{\epsilon^0} = 0. \tag{4.33}$$

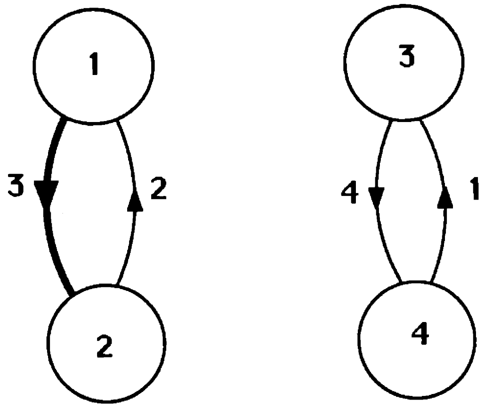


Figure 4.11: Transitions of Interest at Short Time Scale

Therefore the time interval is asymptotically much less than the mean time between the fastest transitions. Hence, if we are in states other than state 1, the expected transition frequency will be 0 (technically it will be $o(\Delta t)$), because more that one transition would have to occur on the interval [t,t+Δt], for a transition from state 1 to state 2 to occur. For state 1, the frequency will be the Markovian transition rate, which in this case is 3 per unit time. This situation is represented diagramatically in Figure 4.11.

Now let us consider a longer time interval. In this case we observe transitions for the interval [t,t+Δt] where Δt = $o(\epsilon^{-1})$ and also require Δt to satisfy

$$\Delta t \succ O(\epsilon^0) = O(1), \tag{4.34}$$

where the symbol $\succ$ indicates that

$$\lim_{\epsilon \to 0} \frac{\epsilon^0}{\Delta t} = 0. \tag{4.35}$$

Therefore, as $\epsilon \to 0$, Δt becomes very large relative to 1 so that a large number of O(1) transitions (in our example the vertical transitions) occur. However, since

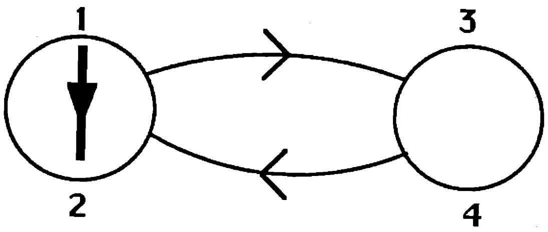


Figure 4.12: Transitions of Interest at the Second Time Scale

$\Delta t = o(\epsilon^{-1})$, horizontal transitions will occur with very low probability. Hence, the state of the system will remain either in the left or right pair of states for the entire interval with probability = 1 - o($\epsilon\Delta t$) = 1 - o(ϵ^0). If the time is spent in the right pair of states, the expected transition frequency will be 0 since neither state 1 or state 2 are on the right. If the state of the system is on the left, then the transition frequency will be non-zero. If we use ergodicity results for Markov chains with a single recurrent class, the expected value of this frequency can be calculated by using steady state conditions for the chain since $\Delta t \succ O(1)$. A diagram illustrating the situation at this time scale is provided in Figure 4.12.

Finally the time increment can be increased once again so that

$$\Delta t \succ O(\epsilon^{-1}) \tag{4.36}$$

or

$$\lim_{\epsilon \to 0} \frac{\epsilon^{-1}}{\Delta t} = 0. \tag{4.37}$$

In this case the length of the time interval is large compared to the mean time between horizontal transitions and therefore those transitions will also reach steady

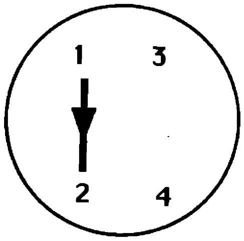


Figure 4.13: Transitions at the Longest Time Scale

state. Therefore, the transition frequency at this time scale should be a weighted combination of the rates for each of the pairs of states with the weights being a function of the fraction of time spent on the left and the fraction of time spent on the right. This situation is depicted in Figure 4.13.

4.5.2 Structured Approach to Calculations

The previous section described the kind of results we can expect for the transition frequencies in a Markov chain with multiple time scales. This section makes precise the problem that we are trying to solve and defines the notation required to provide an algorithmic approach for calculating the transition frequencies and proving the validity of the approach.

We start with the assumption that there is a set of transitions in a Markov chain whose frequency is of interest. Next define the counting process $\eta(t)$ by

$$\eta(t) = \text{the number of transitions of interest up until time t.}$$

In addition, denote the aggregate states at different time scales by $j^{(k)}$, where

$$j^{(k)} \;=\; \text{the j}^{\text{th}} \text{ state obtained in the aggregated chain at time scale k.}$$

Strictly speaking, we should therefore refer to the states in the original unaggregated chain as $1^{(0)}$, $2^{(0)}$, ... , $n^{(0)}$, however the superscript has been dropped at level 0.

In the previous section it was conjectured that there will be different frequencies for each of the aggregate states at any particular time scale. Therefore if the expected value of the frequency is defined, a vector of quantities must be evaluated, with each element corresponding to a particular aggregated state. Therefore, letting ρ(t) be the state of the system at time t, define a vector $\underline{N}^{(k)}(t)$ as the expected transition frequencies at time scale k, with the j^{th} element given by

$$\left[\underline{N}^{(k)}(t)\right]_j = E\left[\eta(t+\Delta t) - \eta(t) \mid O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^k) \text{ and } \rho(t) = j^{(k)}\right] \quad (4.38)$$

and the symbol $\prec$ is defined in Subsection 4.5.1. Note that we may generalize the formulation to the case where more than one set of transition frequencies is determined by defining $\underline{N}^{(k)}(t)$ as a matrix as opposed to a row vector, where each row of the matrix corresponds to an individual set of transitions.

For the example in the previous section, η(t) is the number of transitions from state 1 to state 2. In that example, we considered three possible magnitudes for Δt. If we set k=0, we obtain

$$\left[\underline{N}^{(0)}(t)\right]_j = E\left[\eta(t+\Delta t) - \eta(t) \mid O(\epsilon) \prec \Delta t = o(1) \text{ and } \rho(t) = j^{(0)}\right] \quad (4.39)$$

where $j^{(0)}$ can be 1 to 4 because there are four states in the model at time scale 0. The condition on Δt is given as

$$O(\epsilon) \prec \Delta t = o(1). \quad (4.40)$$

This agrees witht the restriction that Δt be o(1). The additional restriction that $\Delta t \prec O(\epsilon)$ has no effect in this case because there are no faster time scales.

If we set k=1, we obtain

$$\left[\underline{N}^{(1)}(t)\right]_j = E\left[\eta(t+\Delta t) - \eta(t) \mid O(1) \prec \Delta t = o(\epsilon^{-1}) \text{ and } \rho(t) = j^{(1)}\right] \qquad (4.41)$$

This time there are two possible values for $j^{(k)}$, 1 and 2, because there are only two states in the aggregated chain at this time scale. The condition, $O(1) \prec \Delta t = o(\epsilon^{-1})$ matches the condition in Subsection 4.5.1. Finally we may substitute k=2 to obtain an expression similar to (4.41) except that $O(\epsilon^{-1}) \prec \Delta t = o(\ \epsilon^{-2})$. The lower restriction on the magnitude of Δt was introduced in Subsection 4.5.1, while the upper restriction is unimportant as there are no dynamics at longer time scales.

We also define a matrix $\underline{Q}^{(k)}$ according to

$$\underline{Q}^{(k)} = \frac{1}{\Delta t}\underline{N}^{(k)}(t). \qquad (4.42)$$

Having defined the quantities we wish to calculate, two examples are now presented. The transition frequencies are calculated using a technique which is described for a general Markov chain in Subsection 4.5.3.

Example 1:

The Markov chain shown in Figure 4.6 will be used to demonstrate our approach for calculating expected frequencies. Assume that we are interested in the transitions from the top pair of states to the bottom. The first step of the procedure is to form the matrix $\underline{Q}^{(0)}$. This matrix is formed such that $\left[\underline{Q}^{(0)}\right]_j$ is the Markov transition rate out of state j. In our case the transition rates are 3,0,4 and 0 originating in states 1 to 4 respectively. The resulting transition frequency matrix becomes

$$\underline{Q}^{(0)} = \begin{bmatrix} 3 & 0 & 4 & 0 \end{bmatrix}. \qquad (4.43)$$

Recalling that $\underline{U}^{(k-1)}(0)$ is the matrix of ergodic probabilities calculated for the time scale decomposition, the transition frequencies for all of the remaining time scales of the system are calculated using

$$\underline{Q}^{(k)} = \underline{Q}^{(k-1)}\underline{U}^{(k-1)}(0). \tag{4.44}$$

For our example the results are:

$$\underline{Q}^{(1)} = \underline{Q}^{(0)}\underline{U}^{(0)}(0) = \left[\begin{array}{cc} \frac{6}{5} & \frac{4}{5} \end{array}\right] \tag{4.45}$$

and

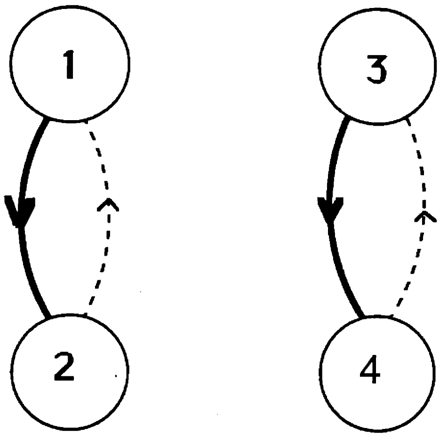
$$\underline{Q}^{(2)} = \underline{Q}^{(1)}\underline{U}^{(1)}(0) = \left[\begin{array}{c} \frac{16}{17}. \end{array}\right] \tag{4.46}$$

Note that $\underline{U}^{(1)}(0)$ was not calculated in Section 4.3, but is easily calculated from (4.17) to obtain

$$\underline{U}^{(1)}(0) = \left[\begin{array}{c} \frac{6}{17} \\ \frac{11}{17} \end{array}\right]. \tag{4.47}$$

The matrix $\underline{Q}^{(1)}$ provides a set of expected frequencies over time scales [t,t+Δt] such that $\Delta t = o(\epsilon^{-1})$. Due to the short time interval under consideration, the system will remain on the left or the right throughout the interval with probability 1 as $\epsilon \to 0$. If the system is on the left, the frequency of transitions from state 1 to state 2 is calculated yielding $\frac{6}{5}$ per unit time. This is obtained by multiplying the rate of 3 from state 1 to state 2 by the ergodic probability (at this time scale) of being in state 1, namely $\frac{2}{5}$. Similarly, if the system is in the right pair of states, the frequency of state 3 to state 4 transitions is determined, which equals $\frac{4}{5}$ per unit time. These rates are indicated in Figure 4.14.

At a longer time scale, such that $t = o(\epsilon^{-2})$, all events are blurred, so that the expected transition frequency becomes a combination of the frequency when the



$$E\left[\frac{\eta_{11}^{(1)}(t)}{t}\right] = \frac{6}{5} \qquad E\left[\frac{\eta_{12}^{(1)}(t)}{t}\right] = \frac{4}{5}$$

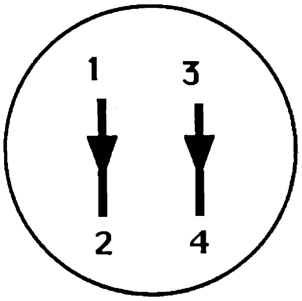
Figure 4.14: Expected Transition Frequencies at Second Time Scale

state of the system is on the left and the frequency when the state of the system is on the right, which yields $\frac{16}{17}$ as given by (4.46). This situation is depicted in Figure 4.15.

Example 2:

In this example, we calculate the expected frequency of a transition that originates in an almost transient state, where we have defined an almost transient state to be a state which is non-transient for $\epsilon \neq 0$, but is transient for $\epsilon = 0$. We will see that this case requires a slightly more complicated procedure. The state transition diagram for this example is shown in Figure 4.16; the only difference from the chain studied in Example 1 is the transition rate from state 2 to state 1.

Suppose that the single transition of interest in this example is that from state 1 to state 2. Omitting the details of the time scale decomposition, the ergodic



$$E\left[\frac{\eta_{11}^{(2)}(t)}{t}\right] = \frac{16}{17}$$

Figure 4.15: Expected Transition Frequencies at Longest Time Scale

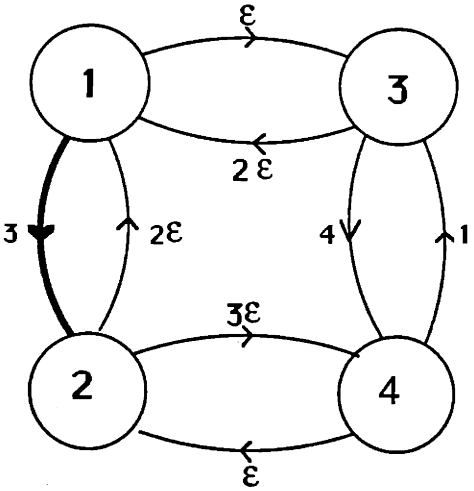


Figure 4.16: State Transition Diagram for Example 2

probabilty matrices are:

$$\underline{U}^{(0)}(0) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & \frac{1}{5} \\ 0 & \frac{4}{5} \end{bmatrix} \tag{4.48}$$

and

$$\underline{U}^{(1)}(0) = \begin{bmatrix} \frac{2}{7} \\ \frac{5}{7} \end{bmatrix}. \tag{4.49}$$

Forming the transition frequency matrix for the fastest time scale we obtain

$$\underline{Q}^{(0)} = \begin{bmatrix} 3 & 0 & 0 & 0 \end{bmatrix}. \tag{4.50}$$

The first element, 3, is just the Markovian transition rate from state 1 to state 2. The remaining elements are zero because the transition of interest originates from state 1 only.

If we calculate $\underline{Q}^{(1)}$ using (4.44) directly, we obtain $\underline{Q}^{(1)} = \underline{0}$, because the ergodic probability of state 1 is 0 in (4.48). In order to avoid this problem, we define a new matrix $\underline{\tilde{Q}}^{(0)}$. This new matrix will be useful when finding $\underline{Q}^{(k)}$ for $k > 0$.

Letting the exact ergodic probability for state i be $\pi_i(\epsilon)$, the problem in this example is that $\pi_1(\epsilon)$ is O(ϵ) and not O(1). To avoid this problem, our first step in finding $\underline{\tilde{Q}}^{(0)}$ is to express $\pi_1(\epsilon)$ as a function of the ergodic probabilities of states in the chain which have O(1) ergodic probabilities. Using the probability mass balance equation at steady state, we may write

$$\pi_1(\epsilon)\ (3 + \epsilon) = \pi_2(\epsilon)\ (2\epsilon) + \pi_3(\epsilon)\ (2\epsilon). \tag{4.51}$$

Solving we obtain

$$\pi_1(\epsilon) = \pi_2(0)\ (\frac{2\epsilon}{3})(1 + O(\epsilon)) + \pi_3(0)\ (\frac{2\epsilon}{3})(1 + O(\epsilon)), \tag{4.52}$$

from which we can obtain the leading order term in the ergodic proabability of state 1. Note that for ϵ=0, the expression reduces to 0 as in (4.48), but since π_2 and π_3 are O(1) (also from (4.40)), the expression for π_1 in (4.52) is non-zero. We can also write

$$3\pi_1(\epsilon) = 0\pi_1(0)\ + 2\epsilon\pi_2(0)(1 + O(\epsilon))\ +\ 2\epsilon\pi_3(0)(1 + O(\epsilon)) + 0\pi_4(0), \tag{4.53}$$

to obtain the transition rate multiplied by the ergodic probability ($\pi_1(\epsilon)$). The quantity in (4.53) is therefore the transition frequency that we are trying to find. Our final step is to assign values to the elements of $\underline{\tilde{Q}}^{(0)}$ according to the leading order coefficients in (4.53) to obtain:

$$\begin{aligned} \left[\underline{\tilde{Q}}^{(0)}\right]_{11} &= 0 \\ \left[\underline{\tilde{Q}}^{(0)}\right]_{12} &= 2\epsilon \\ \left[\underline{\tilde{Q}}^{(0)}\right]_{13} &= 2\epsilon \\ \left[\underline{\tilde{Q}}^{(0)}\right]_{14} &= 0. \end{aligned} \tag{4.54}$$

Combining these results we obtain a matrix of the leading order terms for the expected frequencies or

$$\underline{\tilde{Q}}^{(0)} = \begin{bmatrix} 0 & 2\epsilon & 2\epsilon & 0 \end{bmatrix}. \tag{4.55}$$

We note that states 2 and 3 have O(1) ergodic probabilities, and therefore we may proceed in a similar manner to the first example, but performing the first calculation using

$$\underline{Q}^{(k)} = \underline{\tilde{Q}}^{(k-1)}\ \underline{U}^{(k-1)}(0) \tag{4.56}$$

Therefore we obtain:

$$\begin{aligned}
\underline{Q}^{(1)} &= \tilde{\underline{Q}}^{(0)}\, \underline{U}^{(0)}(0) && (4.57)\\
&= \begin{bmatrix} 2\epsilon & \frac{2\epsilon}{5} \end{bmatrix} && (4.58)\\
\text{and} \qquad \underline{Q}^{(2)} &= \begin{bmatrix} \frac{6\epsilon}{7} \end{bmatrix}. && (4.59)
\end{aligned}$$

In Subsection 4.5.4, we will show that $\underline{Q}^{(k)}$, as calculated by (4.48), yields the leading order terms of the frequencies for transitions of interest, conditioned on the aggregate states defined at the k^{th} time scale. We also show that under certain conditions, the expected frequencies will equal the observed frequencies for a sample path with probability 1.

4.5.3 Algorithm for Event Frequency Calculations

This subsection provides a precise set of steps that can be implemented to obtain the leading order terms of the expected transition frequencies as described in the previous section. The approach starts by performing a time scale decomposition on the Markov chain, which provides both a model of the system at each time scale and the conditional ergodic probability matrices. The algorithm then proceeds to form a matrix of expected frequencies, $\underline{Q}^{(0)}$, for transitions at the shortest time scale. If the states from which the transitions originate are not almost transient states, then the transition frequency matrices for the slower time scales are calculated recursively, using the ergodic probability matrices that we obtained from the time scale decomposition. (An almost transient state is defined as a state which is transient when ϵ=0, but not transient for $\epsilon \neq 0$. These states have ergodic probabilities which are O(ϵ)). If the states from which the transitions originate are almost transient, then a new version of the expected frequency matrix, $\tilde{\underline{Q}}^{(k)}$ is formed and used in place of $\underline{Q}^{(k)}$ to calculate the frequencies at the next time scale. We then proceed as in the O(1) case. The recursive calculations are repeated for each time scale of the system until we reach the time scale at which the entire chain is aggregated into a single state. The algorithm proceeds as follows.

The Algorithm

Before describing the algorithm we define the variables that are introduced by it. First define W_i to be the set of all terminal states for transitions of interest that originate in state i. For example, if we are interested in transitions from state 1 to states 2 or 3 and transitions from state 2 to state 1, then we would have

$$\begin{aligned} W_1 &= \{2,3\} \\ W_2 &= \{1\} \\ W_3 &= \phi \end{aligned} \tag{4.60}$$

where ϕ denotes the empty set. The set $T^{(k)}$ is defined as the set of almost transient states at time scale k, while $M^{(k)}$ is the set of states at time scale k which have O(1) ergodic probability. We also recall that $j^{(k)}$ is the jth state at level k, and $\lambda_{ji}^{(k)}$ is defined as the transition rate from state $i^{(k)}$ to state $j^{(k)}$ at the k^{th} time scale. The superscript (0) is dropped at time scale 0.

Finally let us define expressions for the exact ergodic probability of states as well as approximations to these quantities. We start with $\pi_j(\epsilon)$ which is the exact ergodic probability of state j when all dynamics in the chain have reached steady state. The leading order term of $\pi_j(\epsilon)$ is denoted $\tilde{\pi}_j(\epsilon)$, so that $\pi_j(\epsilon) = \tilde{\pi}_j(\epsilon)(1 + O(\epsilon))$. The quantity $\pi_{j|m}^{(k)}(\epsilon)$ is defined as the conditional ergodic probability of state $j^{(k)}$, given that the chain is in aggregate state $m^{(k+1)}$. It is calculated using the aggregated version of the Markov chain at level k, where the transition rate matrix is assumed to be $\underline{A}^{(k)}(\epsilon)$. Similary $\pi^{(k)}_{j|m}(0)$ is the ergodic probability of state $j^{(k)}$ given aggregate state $m^{(k+1)}$, calculated using $\underline{A}^{(k)}(0)$. Finally, $\tilde{\pi}^{(k)}_{j|m}(\epsilon)$ is the leading order term in $\pi^{(k)}_{j|m}(\epsilon)$ so that

$$\pi^{(k)}_{j|m}(\epsilon) = \tilde{\pi}^{(k)}_{j|m}(\epsilon)(1 + O(\epsilon)). \tag{4.61}$$

Similarly we define a set of conditional ergodic probability expressions for states in the original model at time scale 0, conditioned on aggregate states in the model at any of the K time scales. Let $\pi^{(0,k)}_{j|m}(\epsilon)$ be the ergodic probability of state $j^{(0)}$ (or simply j) given aggregate state $m^{(k)}$. The quantity $\tilde{\pi}^{(0,k)}_{j|m}(\epsilon)$ is defined as the leading order term of $\pi^{(0,k)}_{j|m}(\epsilon)$, while $\pi^{(0,k)}_{j|m}(0)$ is the conditional probabiity calculated for $\epsilon = 0$.

In Theorem 1 that follows this algorithm and the proofs for that theorem, we refer to the magnitude of various quantities in terms of the relationship of the magnitude to the parameter ϵ. Typically, quantities are assumed to have a magnitude that is proportional to some power of ϵ. In order to define magnitude relationships *between* quantities, we use the following notation. If two quantities are related by x=o(y), then

$$\lim_{\epsilon\to 0} \frac{x}{y} = 0. \tag{4.62}$$

We also introduce the notation $x \succ y$ which will be used to indicate that x is lower order in ϵ than y or

$$\lim_{\epsilon\to 0} \frac{y}{x} = 0. \tag{4.63}$$

Finally, we introduce the notation $x \succeq y$ to indicate that x is either lower in order than y or equal in order. Therefore, $x \succeq y$ means that either

$$\lim_{\epsilon\to 0} \frac{y}{x} = 0 \tag{4.64}$$

or

$$\lim_{\epsilon\to 0}\frac{y}{x} = O(1). \tag{4.65}$$

The algorithm that is presented here assumes that there is a single set of transitions whose frequency we are calculating. The case of multiple sets is conceptually identical. In that case there would simply be additional rows added to $\underline{Q}^{(k)}$ and $\underline{\tilde{Q}}^{(k)}$. We may now proceed with the description. Note that we use the symbols $\underline{Q}^{(k)}$ and $\underline{\tilde{Q}}^{(k)}$ to denote the transition frequency matrices without indicating explicit dependence on the parameter ϵ. This is because the matrices will always be functions of ϵ and therefore the argument is dropped for simplicity.

Step 1 : Perform a time scale decomposition on the chain, obtaining descriptions of the dynamics at multiple time scales. By doing this, we obtain the aggregated transition rate matrices, $\underline{A}^{(k)}(\epsilon)$, the ergodic probability matrices, $\underline{U}^{(k)}(0)$, and the leading order terms for the class membership matrices, $\underline{\tilde{V}}^{(k)}(\epsilon)$.

For Example 1 of the previous subsection, this yielded 3 models, described by (4.11) at the fastest time scale. (4.17) at the second time scale, and by $\underline{A}^{(2)}(\epsilon) = [0]$ at the longest time scale (chain becomes a single aggregate state). We also obtained the ergodic probability matrices given by (4.14) and (4.47).

Step 2 : Form the transition frequency matrix $\underline{Q}^{(0)}$ for the shortest time scale according to

$$\left[\underline{Q}^{(0)}\right]_i = \sum_{j\epsilon W_i} \lambda_{ji}. \tag{4.66}$$

To find the elements of the transition frequency matrix for Example 1 of the previous section, we start by defining the sets W_i. We are counting a single transition, from state 1 to state 2, so that

$$W_1 \;=\; \{2\}. \tag{4.67}$$

There are no transitions of interest originating in other states and therefore

$$W_1,\ W_2,\ W_3 \ = \ \phi. \tag{4.68}$$

Using these definitions for the sets W_i, we can use equation (4.62) to calculate

$$\begin{aligned} \left[\underline{Q}^{(0)}\right]_{11} &= \sum_{k \epsilon W_1} \lambda_{k1} \\ &= \lambda_{21} \\ &= 3 \end{aligned} \tag{4.69}$$

and

$$\begin{aligned} \left[\underline{Q}^{(0)}\right]_2, \left[\underline{Q}^{(0)}\right]_3, \left[\underline{Q}^{(0)}\right]_4 &= \sum_{k \epsilon \{\}} \\ &= 0. \end{aligned} \tag{4.70}$$

Step 3: If all of the transitions originate in states which have O(1) ergodic probabilities, then simply define $\underline{\tilde{Q}}^{(k)} = \underline{Q}^{(k)}$. Otherwise the elements of $\underline{\tilde{Q}}^{(k)}$ must be determined from the elements of $\underline{Q}^{(k)}$ by completing the rest of step 3.

3.1 For each state $j^{(k)}$ that satisfies $0 < \pi^{(k)}_{j|m}(\epsilon) \leq O(\epsilon)$ for each $m^{(k+1)}$ and $W_j \neq \phi$ complete step 3.2. Note that $\pi^{(k)}_{j|m}(\epsilon) \leq O(\epsilon)$ if $\pi^{(k)}_{j|m}(0) = 0$, and $\pi^{(k)}_{j|m}(0)$ is simply the corresponding element in $\underline{U}^{(k)}(0)$. (This corresponds to completing step 3.2 if the transitions of interest originate in an almost transient state.)

3.2 Obtain an expression for $\tilde{\pi}^{(k)}_{j|m}(\epsilon)$ by completing the following steps

(a) For each j, so that $\pi^{(k)}_{j|m}(0) = O(1)$, set

$$\tilde{\pi}^{(k)}_{j|m}(\epsilon) = \pi^{(k)}_{j|m}(0) \tag{4.71}$$

for each $m^{(k+1)}$.

(b) For each j of interest, such that $\pi^{(k)}_{j|m}(0) = 0$, but $\pi^{(k)}_{j|m}(\epsilon) \neq 0$, write the expression

$$\tilde{\pi}^{(k)}_{j|m}(\epsilon) = \sum_{i \neq j} \left(\frac{\lambda^{(k)}_{ji} \ \tilde{\pi}^{(k)}_{i|m}(\epsilon)}{\sum_{p \neq j} \ \lambda^{(k)}_{pj}} \right). \tag{4.72}$$

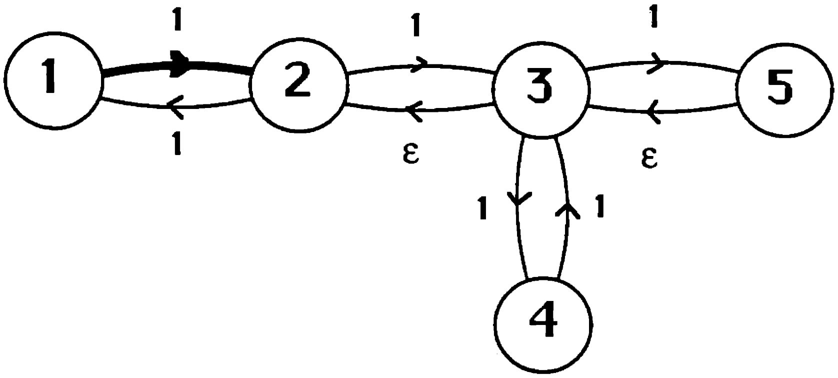


Figure 4.17: Solving for O(ϵ) Ergodic Probabilities

(c) If the right hand side of (4.68) involves other terms $\tilde{\pi}_{i|m}^{(k)}(\epsilon)$ that have not been specified in (a), write (4.68) for these as well.

(d) Repeat step (c) until a closed set of equations is obtained.

The resulting value of $\tilde{\pi}_{j|m}^{(k)}(\epsilon)$ is a linear combination of the ergodic probabilities $\pi_{i|m}^{(k)}(0)$, i$\in M^{(k)}$, and therefore if $b_{ji}^{(k)}(\epsilon)$ are constants obtained from solving the equations, we can write $\tilde{\pi}_{j|m}^{(k)}(\epsilon)$ as

$$\tilde{\pi}_{j|m}^{(k)}(\epsilon) = \sum_{i \in M^{(k)}} b_{ji}^{(k)} \; \pi_{i|m}^{(k)}(0). \qquad (4.73)$$

The elements of $\tilde{\underline{Q}}^{(k)}$ are then calculated as

$$\left[\tilde{\underline{Q}}^{(k)}\right]_i = \begin{cases} \left[\underline{Q}^{(k)}\right]_i + \sum_{j \notin M^{(k)}} b_{ji}^{(k)} \left[\underline{Q}^{(k)}\right]_j , & i \in M^{(k)} \\ 0, & \text{otherwise.} \end{cases} \qquad (4.74)$$

To clarify step 3 of the algorithm consider the chain shown in Figure 4.17 and suppose that we are interested in transitions from state 1 to state 2. We can find

$\underline{Q}^{(0)}$ from step 2 obtaining

$$\underline{Q}^{(0)} = [1\ 0\ 0\ 0\ 0]\,. \tag{4.75}$$

Now, for step 3.1 we see that we are interested in transitions from a state (state 1) that is almost transient. Therefore we proceed to step 3.2 and using equation (4.68) obtain

$$\tilde{\pi}_1(\epsilon) = \tilde{\pi}_2(\epsilon). \tag{4.76}$$

Note that there is no need to condition the ergodic probabilities on individual ergodic classes here, because there is only a single ergodic class for $\epsilon = 0$.

Next we observe that state 2 is also transient at ϵ=0. Therefore we must use (4.68) for the other transient states and we obtain the following set of equations.

$$\begin{aligned} 2\tilde{\pi}_2(\epsilon) &= \tilde{\pi}_1(\epsilon)\ + \epsilon\ \tilde{\pi}_3(\epsilon) \\ \tilde{\pi}_3(\epsilon) &= \tilde{\pi}_4(\epsilon) \\ (2+\epsilon)\ \tilde{\pi}_3(\epsilon) &= \tilde{\pi}_2(\epsilon)\ +\ \tilde{\pi}_4(\epsilon)\ +\ \epsilon\tilde{\pi}_5(\epsilon) \end{aligned} \tag{4.77}$$

We can solve this set of equations and use the fact that $\pi_5(\epsilon)$ is O(1) to obtain

$$\pi_1 = \epsilon^2\ \pi_5(0). \tag{4.78}$$

Finally we obtain (4.75) from step 3.2.

$$\tilde{\underline{Q}}^{(0)} = \left[0\ 0\ 0\ 0\ \epsilon^2\right] \tag{4.79}$$

<u>*Step* 4</u>: Find $\underline{Q}^{(k+1)}$ using

$$\underline{Q}^{(k+1)} = \tilde{\underline{Q}}^{(k)}\ \underline{U}^{(k)}(0). \tag{4.80}$$

This was demonstrated for the second example in the previous subsection (as well as the first example, but with $\tilde{\underline{Q}}^{(k)} = \underline{Q}^{(k)}$.)

Step 5: If there are more time scales, k ← k+1, go to step 3.

The five steps of the algorithm described above will generate a set of matrices $\underline{Q}^{(0)}$, $\underline{Q}^{(1)}$... $\underline{Q}^{(K)}$ for a Markov chain that displays behavior at K different time scales. We may then state the following result which is proved in Subsection 4.5.4.

Theorem 1 : Let $\underline{Q}^{(k)}$ be the vector of expected frequencies defined by the algorithm above, $\left[\underline{Q}^{(k)}\right]_m$ is the m^{th} element of this vector, $\eta(t)$ is the number of transitions of interest that occur by time t, and $\rho(t)$ denotes the Markov chain.

If $O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k})$ then

$$\lim_{\epsilon \to 0} \frac{\left| \left[\underline{Q}^{(k)}\right]_m - E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right] \right|}{\left[\underline{Q}^{(k)}\right]_m} = 0 \qquad (4.81)$$

4.5.4 Proof of Transition Frequency Calculations

Transitions From Recurrent States

We proceed now to prove the validity of the calculations described in the preceding section. We start with the case where states from which transitions are to be counted have O(1) ergodic probabilities. The proof is carried out assuming a single transition in the chain is being counted. Once the proof is complete, an extension is made to cases where a number of different transitions are being considered.

Following the proof for transitions originating in states with O(1) probabilities, the result is generalized for cases where transitions originate in states with ergodic probabilities which are O(ϵ). This generalization is accomplished by determining the leading order terms of the O(ϵ) probabilities.

Before proceding with the proof, we introduce some preliminary concepts and terminology used throughout the proof. First we note that each time we refer to a Markov chain, it is assumed to be a finite-state continous-time Markov chain with n states. The chain will be assumed to possess dynamics on K time scales as defined by Rohlicek [1986]. When an arbitrary time scale is being referred to, we use the lower case k. The Markov chain itself is denoted by $\rho(t)$.

The transition rates in the chain will be denoted λ_{ji} for the transition rate from state i to state j. If the time scale is transformed to a new time variable, the change in the transition rates is indicated by a bar, so λ_{ji} becomes $\bar{\lambda}_{ji}$. Finally if the transition rate is defined for a model of the chain at a time scale $k \neq 0$, the rate is denoted $\lambda_{ji}^{(k)}$, while a state in the model of the system at the kth time scale is denoted $m^{(k)}$. If a state at the fastest time scale (0) is being referred to, then the superscript 0 is dropped. If the state of the system is i, such that i is a member of aggregate state $m^{(k)}$, then we say that $\rho(t) \in m^{(k)}$. Finally, let the number of such states at the kth time scale be n_k.

Next we define some terminology for what we call counting processes. Specifically, suppose that we are interested in counting transitions from state I to state J. Then we define the process $\eta_{JI}(t)$ to be the number of transitions from state I to state J up to time t. We also define a quantity $\eta(t)$ without subscripts which represents the total number of transitions of interest in a chain up to time t. Hence, in the special case where only transitions from state I to state J are being counted we have $\eta(t) = \eta_{JI}(t)$. Finally, we may reference a particular counting process using a single subscript $(\eta_i(t))$. The meaning of the subscript in this case will be defined in the context where the symbol is used.

Throughout the work we refer to ergodic probabilities of states, for which there are several related quantities that must be defined. These quantities were defined prior to the algorithm of Section 4.5.3, and the same meaning is attributed to each variable here. However, the following discussion provides a precise description of each quantity. First we define $\pi_i(\epsilon)$ as the exact ergodic probability obtained by calculating

$$\pi_i(\epsilon) = \left[\lim_{t\to\infty} e^{\underline{A}(\epsilon)t}\underline{x}(0)\right]_i, \tag{4.82}$$

that is the ith element of the limiting probability vector as $t \to \infty$. Now suppose we wish to focus on the behavior at a particular, specifically the kth time scale. Let $\pi_{i|m}^{(k)}(\epsilon)$ denote the corresponding ergodic probability at this time scale of being in state i$^{(k)}$ given that the process is in $m^{(k+1)}$. That is

$$\pi_{i|m}^{(k)}(\epsilon) = \lim_{\substack{\epsilon\to 0 \\ O(\epsilon^{-k})\prec\Delta t=o(\epsilon^{-k-1})}} \left[\Pr(\rho(t+\Delta t) \in i^{(k)}) \mid \rho(t) \in m^{(k+1)}\right] \tag{4.83}$$

We may also define some variables related to this quantity. The first such quantity is $\tilde{\pi}_{i|m}^{(k)}(\epsilon)$, which is the leading order term in $\pi_{i|m}^{(k)}(\epsilon)$ such that

$$\pi_{i|m}^{(k)}(\epsilon) = \tilde{\pi}_{i|m}^{(k)}(\epsilon)(1 + O(\epsilon)). \tag{4.84}$$

The second related quantity is the probability that is calculated with $\epsilon = 0$, which is denoted by $\pi_{i|m}^{(k)}(0)$ and is evaluated by setting $\epsilon = 0$ in the expression $\pi_{i|m}^{(k)}(\epsilon)$. Finally, we define $\pi_{i|m}^{(0,k)}(\epsilon)$ to be the ergodic probability of state i in the <u>original</u> model with transition rate matrix $\underline{A}^{(0)}(\epsilon)$ given that we are in aggregate state $m^{(k)}$. Therefore equation (4.83) becomes

$$\pi_{i|m}^{(0,k)}(\epsilon) = \lim_{\substack{\epsilon\to 0 \\ O(\epsilon^{-k+1})\prec\Delta t=o(\epsilon^{-k})}} \left[\Pr(\rho(t+\Delta t) = i) \mid \rho(t) \in m^{(k)}\right] \tag{4.85}$$

The last expressions we describe are those related to the matrices $\underline{U}^{(k)}(\epsilon)$ and $\underline{V}^{(k)}(\epsilon)$. These matrices were described in Section 4.3 as the ergodic probability and class membership matrices. In addition, the corresponding matrices $\underline{U}^{(k)}(0)$ and $\underline{V}^{(k)}(0)$ were defined as the corresponding matrices calculated with $\epsilon = 0$. Now we also define $\tilde{\underline{U}}^{(k)}(\epsilon)$ and $\tilde{\underline{V}}^{(k)}(\epsilon)$ as the matrices of leading order terms of the elements in $\underline{U}^{(k)}(\epsilon)$ and $\underline{V}^{(k)}(\epsilon)$. Therefore we have for the ergodic probability matrix that

$$\left[\underline{U}^{(k)}(\epsilon)\right]_{ij} = \left[\underline{\tilde{U}}^{(k)}(\epsilon)\right]_{ij}(1+O(\epsilon)) \tag{4.86}$$

for each i and j. A similar result holds for the class membership matrix.

In what follows, as in (4.77), we will be considering time intervals with lengths that are different orders of ϵ. For example, in Lemma 1, we consider $\Delta t = o(1)$, i.e. we are explicitly considering Δt to be a function of ϵ that goes to zero as $\epsilon \rightarrow 0$.

Now we may proceed with the proof. Some minor results required to prove the main theorem are provided first. Lemma 1 starts by providing a result which shows that for short time intervals, the probability of more than 1 transition is small enough so that we may perform all calculations assuming that at most one transition may occur.

Lemma 1

Suppose that we have an n-state Markov chain and we are interested in the transitions from state I to state J, and that $\lambda_{JI} = \mathrm{O}(1)$. Since we are counting only one transition we have that $\eta(t) = \eta_{JI}(t)$. Then if

$$P_N = Pr\{\eta(t+\Delta t) - \eta(t) > N\}. \tag{4.87}$$

Then for $\Delta\mathrm{t} = \mathrm{o}(1)$ we have

$$\lim_{\epsilon \to 0} \frac{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right] - \frac{P_1}{\Delta t}}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]} = 0 \tag{4.88}$$

Proof:

Since P_N is the probability of at least N transitions, we know that

$$\begin{aligned} E\left[\eta(t+\Delta t) - \eta(t)\right] &= 1(P_1 - P_2) + 2(P_2 - P_3)... \\ &= P_1 + P_2 + \dots . \end{aligned} \tag{4.89}$$

Now consider a second chain obtained from the first one by removing all transitions out of state I except the transition to state J. Then if we let P'_N denote the corresponding probabilities for this second chain, we obviously have that

$$P_N \leq P'_N. \tag{4.90}$$

However,

$$\begin{aligned} P'_1 &= \int_0^{\Delta t} \lambda_{JI}\, e^{-\lambda_{JI} t}\, dt \\ &= 1 - e^{-\lambda_{JI}\Delta t}. \end{aligned} \tag{4.91}$$

Now, using the fact that λ_{JI} is O(1) and Δt is o(1), we know that $\lambda_{JI}\ \Delta t$ is small with respect to 1. Combining this with the fact that $P_1 \leq P'_1$, we obtain

$$P_1 \leq \lambda_{JI}\ \Delta t. \tag{4.92}$$

The next step is to obtain a bound on the magnitude of P_2. To do this we first write

$$P_2 < \int_0^{\Delta t} Pr\{\text{Trans. at t} = \tau\}\ \text{Pr}\{\text{Trans. during } [\tau, \Delta\text{t}]\} \text{d}\tau \tag{4.93}$$

because two transitions would require an I to J transition followed by transitions back to state I and then a second I to J transition. Therefore

$$\begin{aligned} P_2 &< \int_0^{\Delta t} \lambda_{JI} e^{-\lambda_{JI}\tau} d\tau \left(\int_0^{\Delta t - \tau} \lambda_{JI} e^{-\lambda_{JI} t} dt \right) \\ &\leq \int_0^{\Delta t} \lambda_{JI} e^{-\lambda_{JI}\tau} d\tau\ P_1 \\ &\leq P_1^2 \end{aligned} \tag{4.94}$$

Similarly we can easily show that $P_N \leq P_1^N$. Therefore

$$E\left[\eta(t + \Delta t) - \eta(t)\right] = P_1 + P_2 + ...$$

$$\begin{aligned} &= P_1\left[1 + \frac{P_2}{P_1} + \frac{P_3}{P_1} + ...\right] \\ &\leq P_1\left[1 + P_1 + P_1^2\cdots\right] \\ &= \frac{P_1}{(1-P_1)}. \end{aligned} \qquad (4.95)$$

Also, we obviously have that

$$E\left[\eta(t+\Delta t) - \eta(t)\right] > P_1. \qquad (4.96)$$

Therefore

$$\frac{P_1}{\Delta t} < E\left[\frac{\eta(t+\Delta t) - \eta(t)}{\Delta t}\right] < \frac{P_1}{(1-P_1)\Delta t} \qquad (4.97)$$

so

$$\begin{aligned} 0 < \frac{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right] - \frac{P_1}{\Delta t}}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]} &< \frac{P_1^2}{(1-P_1)\Delta t E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]} \\ &< \frac{P_1^2\Delta t}{(1-P_1)\Delta t P_1} \\ &= \frac{P_1}{1-P_1} \end{aligned} \qquad (4.98)$$

However, since $\Delta t = o(1)$,

$$\lim_{\epsilon \to 0} \frac{P_1}{1-P_1} = 0 \qquad (4.99)$$

and therefore our result is proved.

Corollary

For $\lambda_{JI} = \text{o}(1)$ the result of Lemma 1 also holds.

Proof:

The same proof holds only now we have that $P_1 \leq o(1)^2$.

Now we proceed to prove two lemmas related to the calculation of the expected values of the frequency of a single transition. Expressions are obtained which are functions of both the probability of the state of origin of the transition of interest and the rate for the transition of interest. We start with Lemma 2 which provides a result for short time intervals.

Lemma 2

Suppose that we have the same Markov chain as in Lemma 1 and that the same variable definitions and assumptions apply. In addition assume that $\Delta t = o(1)$. Then the following is true:

$$\lim_{\epsilon \to 0} \frac{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right] - \frac{1}{\Delta t}\int_t^{t+\Delta t} \lambda_{JI} Pr(\rho(\tau) = I)\, d\tau}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]} = 0 \qquad (4.100)$$

<u>Proof</u>:

The first thing that we do is to replace the original chain with one that is equivalent for our purposes. From Lemma 1, we can calculate the desired expected value from the probability that at least one transition from state I to state J has occurred. Therefore, any transitions on the interval $[t, t+\Delta t]$ which follow a transition from state I to state J are irrelevant and therefore, we can replace the transition from state I to state J with a transition to a new state J' such that J' is a trapping state. Therefore $\lambda_{iJ'} = 0$ for all i, where i is an arbitrary state in the chain.

We now draw on a result from Keilson [16]. Keilson deals with an n-state chain of the form we are discussing and defines a process T(t) such that T(t) is an ordered pair representing the starting and ending states of the last transition. Hence if T(t) = (i,j), then the last transition prior to time t was from state i to state j. His result, applied to our particular transition pair (I, J') is

$$\frac{d}{dt}\, Pr\{T(t) = (I, J')\} = -\sum_{\substack{i \neq J' \\ i=1}}^{n} \lambda_{iJ'}\, Pr\{\rho(t) = J'\} + \lambda_{J'I}\, Pr\{\rho(t) = I\}. \qquad (4.101)$$

However, having replaced the original chain by the chain with trapping state J', the left-hand side is just the time derivative with respect to τ of $\Pr(\eta(\tau) > \eta(t))$. Then, letting $\epsilon \rightarrow 0$ we can revert back to the original chain, showing that

$$\lim_{\epsilon\rightarrow 0} \frac{\frac{d}{d\tau}\ Pr\{\eta(\tau) > \eta(t)\} - \lambda_{JI}\ Pr(\rho(\tau) = I)}{\frac{d}{d\tau}\ Pr\{\eta(\tau) > \eta(t)\}} = 0. \qquad (4.102)$$

Hence we have

$$\lim_{\epsilon\rightarrow 0} \frac{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right] - \frac{1}{\Delta t}\int_t^{t+\Delta t} \lambda_{JI} Pr\{\rho(\tau) = I\}\ d\tau}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]} = 0 \qquad (4.103)$$

and the result is proved.

Lemma 3 provides an analogous result for long time intervals.

Lemma 3

Suppose we have a Markov chain with a single ergodic class for $\epsilon > 0$ and we are interested in the frequency of transitions from state I to state J. Now if we have $\lambda_{JI} \succeq O(1)$, and $\pi_{I|m}^{(0,k)} = \mathrm{O}(1)$ for some aggregate state $m^{(k)}$ and $\Delta t \succ O(\epsilon^{-k+1})$, then

$$\lim_{\epsilon\rightarrow 0} \frac{\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right] - \lambda_{JI}\pi_{I|m}^{(0,k)}}{\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right]} = 0 \qquad (4.104)$$

<u>Proof</u>:

The proof of the lemma closely follows the proof on pages 230 to 232 of [24] for what Ross refers to as the *elementary renewal theorem.* To simplify our notation we will work with t and 0 as opposed to t+Δt and t. The proof is unaffected. Introduce a new process ν_N such that ν_N is the time at which the Nth transition from state I to state J occurs. Also, let X_N be the time between transition N-1 and transition N. Now scale time so that $\tau = \epsilon^{k-1}t$ yielding $\tau \succ \mathrm{O}(1)$, while the scaled transition rate satisfies $\bar{\lambda}_{JI} \succeq O(\epsilon^{-k+1}) \succeq \mathrm{O}(1)$. Therefore, we have that

$$\nu_{\eta(\tau)} \leq \tau \leq \nu_{\eta(\tau)+1}, \qquad (4.105)$$

and dividing by $\eta(\tau)$ we obtain

$$\frac{\nu_{\eta(\tau)}}{\eta(\tau)} \leq \frac{\tau}{\eta(\tau)} < \frac{\nu_{\eta(\tau)+1}}{\eta(\tau)}. \tag{4.106}$$

Now since the chain is assumed to be positive recurrent,

$$\Pr\{\nu_{\eta(\tau)+1} - \nu_{\eta(\tau)} < \infty\} = 1. \tag{4.107}$$

It follows that

$$\Pr\{\nu_{\eta(\tau)} < \infty\} = 1 \tag{4.108}$$

for finite $\eta(\tau)$ and

$$\lim_{\tau\to\infty} \eta(\tau) = \infty. \tag{4.109}$$

Now, from the strong law of large numbers ([24], page 231) and (4.109),

$$\lim_{\tau\to\infty} \frac{\nu_{\eta(\tau)}}{\eta(\tau)} = \lim_{\eta(\tau)\to\infty} \frac{\nu_{\eta(\tau)}}{\eta(\tau)} = \bar{X},\ w.p.1. \tag{4.110}$$

where $\bar{X}$ is the mean time between transitions from state I to state J. Similarly, using the fact that $\Pr\{X_{N+1} = \infty\} = 0$, we find that

$$\begin{aligned} \lim_{\tau\to\infty} \frac{\nu_{\eta(\tau)+1}}{\eta(\tau)} &= \lim_{\eta(\tau)\to\infty} \frac{\nu_{\eta(\tau)+1}}{\eta(\tau)} \\ &= \lim_{\eta(\tau)\to\infty} \frac{\nu_{\eta(\tau)}}{\eta(\tau)} + \lim_{\eta(\tau)\to\infty} \frac{X_{N+1}}{\eta(\tau)} \\ &= \bar{X}. \end{aligned} \tag{4.111}$$

From the fact that $\tau \succ O(1)$ and equations (4.106),(4.110) and (4.111), we now have that

$$\lim_{\epsilon\to 0} \frac{\eta(\tau)}{\tau} = \lim_{\tau\to\infty} \frac{\eta(\tau)}{\tau} = \frac{1}{\bar{X}}. \tag{4.112}$$

It remains to calculate $\bar{X}$. However, from Kielson [1980], for a chain with one ergodic class,

$$\bar{X} = \lambda_{JI}\pi_I(\epsilon) = \lambda_{JI} \lim_{t\to\infty} \Pr\{\rho(t) = I\}, \tag{4.113}$$

but

$$\begin{aligned} \pi_{I|m}^{(0,k)} &= \lim_{\substack{\epsilon\to 0 \\ O(\epsilon^{-k+1})\prec t=o(\epsilon^{-k})}} \Pr\{\rho(\tau)=I \mid \rho(0)\in m^{(k)}\} \\ &= \lim_{\tau\to\infty} \Pr\{\rho(\tau)=I \mid \rho(0)\in m^{(k)}\} \qquad (4.114) \end{aligned}$$

and therefore

$$\frac{1}{\bar{X}} = \lambda_{JI}\pi_{I|m}^{(0,k)}. \qquad (4.115)$$

We therefore have the desired result from (4.112) and (4.115) that (transforming back to Δt),

$$\lim_{\epsilon\to 0}\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t)\in m^{(k)}\right] = \lambda_{JI}\pi_{I|m}^{(0,k)}. \qquad (4.116)$$

Now, using the fact that $\lambda_{JI} \succeq \mathrm{O}(1)$ and $\pi_{I|m}^{(0,k)} = \mathrm{O}(1)$,

$$\lim_{\epsilon\to 0}\frac{\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t)\in m^{(k)}\right] - \lambda_{JI}\pi_{I|m}^{(0,k)}}{\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t)\in m^{(k)}\right]} = 0. \qquad (4.117)$$

Next we prove a result related to the ergodic probability of a state I conditioned on each aggregate state $m^{(k)}$. Suppose that we have a Markov chain that exhibits behavior at K different time scales (in the sense described in Section 4.3). Let $\underline{x}(t)$ be the vector of state probabilities at time t, while $\underline{U}^{(k)}(0)$, $\tilde{\underline{V}}^{(k)}(\epsilon)$, and $\underline{A}^{(k)}(\epsilon)$ are the ergodic probability, class membership and transition rate matrices as described in Section 4.3. Note that $\tilde{\underline{V}}^{(k)}(\epsilon)$ is a function of ϵ, but contains only the leading order terms of the exact class membership matrix. Also let $\underline{x}^{(k)}(t)$ be vectors of probabilities at each time scale such that $\underline{\mathrm{x}}(0)$ is the vector of the probabilities of the states at t=0 and

$$\begin{aligned} \underline{x}^{(0)}(t) &= \underline{x}(t) = e^{\underline{A}^{(0)}(\epsilon)t}\underline{x}(0) \\ \underline{x}^{(k)}(t) &= \tilde{\underline{V}}^{(k-1)}(\epsilon)\underline{x}^{(k-1)}(t). \end{aligned} \tag{4.118}$$

Finally let $\underline{q}^{(k)}$ be a row vector defined by

$$\begin{aligned} \left[\underline{q}^{(0)}\right]_i &= \begin{cases} 1 & i = I \\ 0 & \text{otherwise} \end{cases} \\ \left[\underline{q}^{(k)}\right] &= \underline{q}^{(k-1)}\ \underline{U}^{(k-1)}(0). \end{aligned} \tag{4.119}$$

Then we have the following result:

Lemma 4:

If Δt satisfies the constraint

$$O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k}) \tag{4.120}$$

then

$$\lim_{\substack{\epsilon \to 0 \\ O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k})}} Pr\{\rho(t) = I\} = \underline{q}^{(k)}\ \underline{x}^{(k)}(t) + O(\epsilon) \tag{4.121}$$

Proof:

If we let $\underline{A}^{(0)}(\epsilon)$ be the transition rate matrix for the Markov chain, then the results from Rohlicek [1986] indicate that we can express the state probability vector as

$$\begin{aligned} \underline{x}(t+\Delta t) &= \Big(e^{\underline{A}^{(0)}(\epsilon)\Delta t} - \underline{U}^{(0)}\underline{\tilde{V}}^{(0)}(\epsilon) \\ &+ \underline{U}^{(0)} e^{\underline{A}^{(1)}(\epsilon)\epsilon\Delta t}\underline{\tilde{V}}^{(0)}(\epsilon) - \underline{U}^{(0)}\underline{U}^{(1)}\underline{\tilde{V}}^{(1)}(\epsilon)\underline{\tilde{V}}^{(0)}(\epsilon) \\ &\vdots \\ &+ \underline{U}^{(0)}\cdots\underline{U}^{(K-1)} e^{\underline{A}^{(K)}(\epsilon)\epsilon^{K}\Delta t}\underline{\tilde{V}}^{(K-1)}(\epsilon)\cdots\underline{\tilde{V}}^{(0)}(\epsilon) \\ &- \underline{U}^{(0)}\cdots\underline{U}^{(K-2)}\underline{\tilde{V}}^{(K-2)}(\epsilon)\cdots\underline{\tilde{V}}^{(0)}(\epsilon)\Big)\,\underline{x}(t) + O(\epsilon). \qquad (4.122) \end{aligned}$$

If we consider the special case of

$$O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k}), \qquad (4.123)$$

we obtain

$$\lim_{\substack{\epsilon\to 0 \\ O(\epsilon^{-k+1})\prec\Delta t=o(\epsilon^{-k})}} \underline{x}(t+\Delta t) = \underline{U}^{(0)}\cdots\underline{U}^{(k-1)}\underline{\tilde{V}}^{(k-1)}(\epsilon)\cdots\underline{\tilde{V}}^{(0)}(\epsilon)\underline{x}(t) + O(\epsilon). \qquad (4.124)$$

This will yield an nx1 vector of probabilities for the n states. In the statement that we are proving, we are interested in the probability of state I. To obtain this probability we premultiply by a row vector $\underline{q}^{(0)}$ such that

$$\left[\underline{q}^{(0)}\right]_i = \begin{cases} 1, & i = I \\ 0, & \text{otherwise} \end{cases} \qquad (4.125)$$

and using the definitions for $\underline{q}^{(k)}$ and $\underline{x}^{(k)}(t)$ given in the statement of Lemma 4 we can substitute to obtain

$$\begin{aligned} \lim_{\substack{\epsilon\to 0 \\ O(\epsilon^{-k+1})\prec\Delta t=o(\epsilon^{-k})}} Pr\{\rho(t+\Delta t)=I\} &= \underline{q}^{(0)}\underline{U}^{(0)}\cdots\underline{U}^{(k-1)}\underline{V}^{(k-1)}(\epsilon)\cdots\underline{V}^{(0)}(\epsilon)\underline{x}(t) + O(\epsilon) \\ &= \underline{q}^{(k)}\underline{x}^{(k)}(t) + O(\epsilon) \qquad (4.126) \end{aligned}$$

as required.

We now proceed with Theorem 2 which provides a technique for calculating transition frequencies when the ergodic probability of the states from which the transitions originate are O(1).

Consider a Markov chain with transition matrix $\underline{A}^{(0)}(\epsilon)$. The matrices $\underline{U}^{(k)}(0)$ and $\underline{V}^{(k)}(\epsilon)$ are defined in Section 4.3. We now state a theorem for the calculation of transition frequencies when the state of origin has an O(1) ergodic probability. The following theorem is stated for the special case where we are interested in only a single transition. The result is generalized following the proof of the theorem.

Theorem 2

Suppose that we are interested in calculating the frequency of a transition from state I to state J. Therefore we define a counting process $\eta(t)$ that equals the number of transitions from state I to state J up to time t. Furthermore, let $\pi^{(0,k)}_{I|m}(\epsilon)$ be O(1) or zero for each $m^{(k)}$. Define $\underline{Q}^{(k)}$ by

$$\underline{Q}^{(k)} = \lambda_{JI}\underline{q}^{(k)}, \tag{4.127}$$

where $\underline{q}^{(k)}$ and $\underline{x}^{(k)}(t)$ have the same definitions as in Lemma 4. If we let

$$O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k}) \tag{4.128}$$

then

$$\lim_{\epsilon \to 0} \frac{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right] - \left[\underline{Q}^{(k)}\right]_m}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right]} = 0 \tag{4.129}$$

Proof:

We start by letting the transition rate λ_{JI} be O(ϵ^p). There are two cases to consider, p $\leq$ k-1 and p $\geq$k. Since the order of p is restricted to be an integer power of ϵ, these are the only two possibilities. Start with Case 1.

Case 1: p $\leq$ k-1

In this case we start by scaling time, with the new time variable τ given by

$$\tau = \epsilon^{k-1}t. \tag{4.130}$$

Therefore,

$$O(1) \prec \Delta\tau = o(\epsilon^{-1}), \tag{4.131}$$

while the new transition rate symbolized by $\bar{\lambda}_{JI}$ satisfies

$$\begin{aligned} \bar{\lambda}_{JI} &= \frac{\lambda_{JI}}{\epsilon^{k-1}} \\ &= O(\epsilon^{p-k+1}) \\ &\succeq O(1), \end{aligned} \tag{4.132}$$

and therefore the conditions of Lemma 3 apply. Using the results of Lemma 3 and then rescaling the time variable, we obtain

$$\lim_{\epsilon\to 0} \frac{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right] - \lambda_{JI}\pi^{(0,k)}_{I|m}(\epsilon)}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right]} = 0, \tag{4.133}$$

which, using the definition of $\pi^{(k)}_{I|m}(\epsilon)$ and $\underline{x}^{(k)}(t)$, as well as the results of Lemma 4 yields

$$\begin{aligned} \lim_{\epsilon\to 0} \frac{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right] - \underline{Q}^{(k)}\underline{x}^{(k)}(t)}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]} &= \lim_{\epsilon\to 0} O(\epsilon) \\ &= 0 \end{aligned} \tag{4.134}$$

as required.

Case 2: p ≥ k

The proof for this case is similar to the previous case except for the scaling and the particular Lemmas that are applied. We start by scaling the time variable so that

$$\tau = \epsilon^k t \tag{4.135}$$

obtaining

$$\bar{\lambda}_{JI} \preceq O(1) \tag{4.136}$$

and

$$O(\epsilon) \ < \ \Delta\tau \ = o(1). \tag{4.137}$$

Applying Lemmas 2 and 4 we have that

$$\begin{aligned} 0 &= \lim_{\epsilon \to 0} \frac{E\left[\frac{\eta(\tau+\Delta\tau)-\eta(\tau)}{\Delta\tau}\right] - \int_{\tau}^{\tau+\Delta\tau} \frac{1}{\Delta\tau}\lambda_{JI} Pr\{\rho(\tau') = I\} d\tau'}{E\left[\frac{\eta(\tau+\Delta\tau)-\eta(\tau)}{\Delta\tau}\right]} \\ &= \lim_{\epsilon \to 0} \frac{E\left[\frac{\eta(\tau+\Delta\tau)-\eta(\tau)}{\Delta\tau}\right] - \int_{\tau}^{\tau+\Delta\tau} \frac{1}{\Delta\tau}\lambda_{JI} Pr\{\rho(\tau+\Delta\tau) = I\} d\tau'}{E\left[\frac{\eta(\tau+\Delta\tau)-\eta(\tau)}{\Delta\tau}\right]} \\ &= \lim_{\epsilon \to 0} \frac{E\left[\frac{\eta(\tau+\Delta\tau)-\eta(\tau)}{\Delta\tau}\right] - \underline{Q}^{(k)}\underline{x}^{(k)}(\tau)}{E\left[\frac{\eta(\tau+\Delta\tau)-\eta(\tau)}{\Delta\tau}\right]}. \end{aligned} \tag{4.138}$$

Combining cases 1 and 2 we have that equation (4.138) holds in general. Now if we have that $\rho(t) \in m^{(k)}$ then

$$\left[\underline{x}^{(k)}(t)\right]_i = \begin{cases} 1, & i = m \\ 0, & i \neq m \end{cases} \tag{4.139}$$

and hence

$$\left[\underline{Q}^{(k)}\underline{x}^{(k)}(t)\right] = \left[\underline{Q}^{(k)}\right]_m . \tag{4.140}$$

Therefore the theorem is proved.

Corollary

If we have the conditions of Theorem 2 and are counting N individual transitions, represented by counting processes $\eta_1(t)$, $\eta_2(t)$,...,$\eta_N(t)$ and we define $\underline{Q}_i^{(k)}$ for each count process as we did in the statement of Theorem 2 for one count process, then we may obtain the matrix $\underline{Q}^{(k)}$ if we are counting all of the transitions collectively, where

$$\underline{Q}^{(k)} = \sum_{i=1}^{N} \underline{Q}_i^{(k)}. \tag{4.141}$$

Proof If $\eta(t)$ is the total number of transitions that occur, from the mutual exclusivity of transitions,

$$\eta(t) = \sum_{i=1}^{N} \eta_i(t). \tag{4.142}$$

Hence we may show that

$$\begin{aligned}
\lim_{\epsilon\to 0}\left[\frac{\underline{Q}^{(k)}\underline{x}^{(k)}(t) - E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]}\right] &= \lim_{\epsilon\to 0}\left[\frac{\sum_{i=1}^{N}\underline{Q}_i^{(k)}\underline{x}^{(k)}(t) - \sum_{i=1}^{N} E\left[\frac{\eta_i(t+\Delta t)-\eta_i(t)}{\Delta t}\right]}{\sum_{i=1}^{N} E\left[\frac{\eta_i(t+\Delta t)-\eta_i(t)}{\Delta t}\right]}\right] \\
&\leq \lim_{\epsilon\to 0}\sum_{i=1}^{N}\left[\frac{\underline{Q}_i^{(k)}\underline{x}^{(k)}(t) - E\left[\frac{\eta_i(t+\Delta t)-\eta_i(t)}{\Delta t}\right]}{E\left[\frac{\eta_i(t+\Delta t)-\eta_i(t)}{\Delta t}\right]}\right] \\
&= 0.
\end{aligned} \tag{4.143}$$

and the result is proved.

Theorem 2 provides a formula for the expected transition frequency of a set of transitions, given a particular time horizon $(t, t+\Delta t)$ and aggregate state at time t.

We now use this result to show that the algorithm of Section 4.5.3 yields the desired results. We proceed by constructing $\underline{Q}^{(k)}$ as defined in Theorem 2 according to the following steps.

Step 1: Since we require the ergodic probability and class membership matrices we start by performing a time scale decomposition of the chain.

Step 2: Form the matrix $\underset{\sim}{Q}^{(0)}$. From the corollary to Theorem 2,

$$\underset{\sim}{Q}^{(0)} = \sum_{i=1}^{N} \underset{\sim}{Q}_i^{(0)} \tag{4.144}$$

where $\underset{\sim}{Q}_i^{(0)}$ are the transition frequency matrices for the individual transitions defined by

$$\left[\underset{\sim}{Q}_i^{(0)}\right]_j = \begin{cases} \lambda_{JI}, & j = I \\ 0, & j \neq I \end{cases} \tag{4.145}$$

if the ith transition of interest is from state I to state J. Therefore, if we define W_I to be the set of all terminal states for transitions of interest that originate in state I, then

$$\left[\underset{\sim}{Q}^{(0)}\right]_I = \sum_{j \in W_I} \lambda_{jI}. \tag{4.146}$$

Step 3: Ensure that all transitions originate in states with O(1) ergodic probability (so that Theorem 2 applies). To conform to the steps in the algorithm we directly assign $\tilde{\underset{\sim}{Q}}^{(k)} = \underset{\sim}{Q}^{(k)}$.

Step 4: Given the definition of $\underset{\sim}{Q}^{(k)}$ in (4.125) and (4.127), we can find $\underset{\sim}{Q}^{(k)} = \underset{\sim}{Q}^{(k-1)}\underset{\sim}{U}^{(k-1)}(0)$.

Step 5: Repeat step 4 for each of the K time scales. Completing steps 1 through 5 we obtain $\underset{\sim}{Q}^{(k)}$ that satisfies Theorem 1. However, comparing these steps to those in the algorithm of Section 4.5.3 in the case where transitions originate in states with O(1) ergodic probability, they are identical. Hence the algorithm is valid for these cases.

Transitions Originating in Almost Transient States

All of the work thus far in proving the validity of our approach to calculating transition frequencies has used the assumption that the state from which the transitions originate are not almost transient. In the case of almost transient states, the straightforward approach considered in the previous section fails for two reasons. First we note that the expression for probabilities quoted from Rohlicek [1986] is of the form

$$\underline{x}(t) = [\cdots] \; + \; O(\epsilon), \tag{4.147}$$

and therefore calculates only the O(1) terms of the state probabilities. If (4.147) is used for states which have O(ϵ) probabilities or smaller, a result of 0 will be obtained. This is inadequate because the state probability is not zero unless ϵ=0 and therefore the expected transition frequency is not 0.

The second problem is created by Lemma 3 where we also assumed that the ergodic probabilities of the states from which the transitions originate are O(1). If this is not the case, the Lemma does not apply unless further contraints are placed on Δt. This is because the proof relies on the fact that many of the transitions of interest occur over the interval $[t, t + \Delta t]$. If the ergodic probability of the state of origin is small enough this will not be the case.

We proceed to modify the approach and prove the validity of our modification by taking the following steps;

1) Demonstrating an example which fails because the transition of interest originates in an almost transient state and showing that our algorithm corrects the problem.

2) Showing that we would obtain the correct result if we knew the leading order terms of the ergodic probabilities for all states from which transitions originate

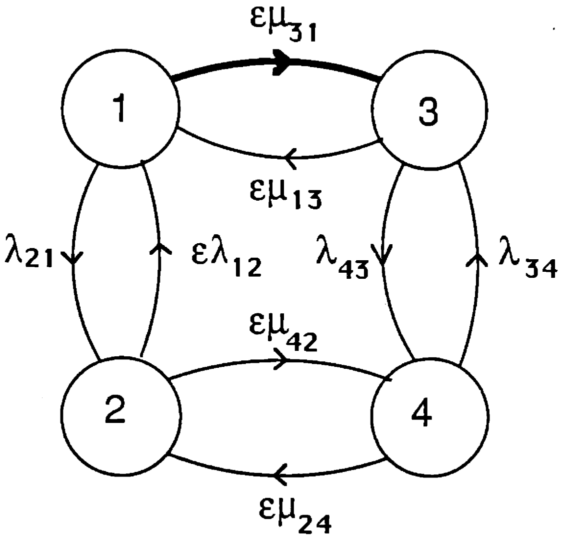


Figure 4.18: Multiple Time Scales and Transient States

regardless of the magnitude of that leading term.

3) Showing that the steps taken in step 3.2 of our algorithm in Section 4.5.3 are equivalent to calculating the leading order terms for the ergodic proabablities of the states from which the transitions of interest originate. Combining this result with step 2 above, the validity of the algorithm is proven.

We start with an example in which transitions originating in an almost transient state are being counted.

Example:

Suppose that we have the system of Figure 4.18 and that we are interested in counting the transitions from state 1 to state 3.

We may form $\underline{Q}^{(0)}$ based on (4.62) to obtain:

$$\underline{Q}^{(0)} = [\epsilon\mu_{31}\ 0\ 0\ 0]\,. \tag{4.148}$$

We may also calculate the ergodic probability matrices evaluated at $\epsilon = 0$ obtaining,

$$\underline{U}^{(0)}(0) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & \frac{\lambda_{34}}{\lambda_{43}+\lambda_{34}} \\ 0 & \frac{\lambda_{43}}{\lambda_{43}+\lambda_{34}} \end{bmatrix} \tag{4.149}$$

and

$$\underline{U}^{(1)}(0) \;=\; \begin{bmatrix} \frac{\lambda_{34}\mu_{42}+\lambda_{43}\mu_{24}}{\lambda_{34}\mu_{42}+\lambda_{43}\mu_{24}+\mu_{13}(\lambda_{43}+\lambda_{34})} \\ \frac{\mu_{13}(\lambda_{43}+\lambda_{34})}{\lambda_{34}\mu_{42}+\lambda_{43}\mu_{24}+\mu_{13}(\lambda_{4}3+\lambda_{34})} \end{bmatrix} \tag{4.150}$$

If we attempt to calculate $\underline{Q}^{(1)}$ directly, we obtain

$$\underline{Q}^{(1)} = [0 \; 0 \,], \tag{4.151}$$

which is incorrect because it yields only frequencies of 0. However the transitions do take place and therefore must have a non-zero expected frequency. Our algorithm solves this problem by calculating the leading order terms of the ergodic probabilities. To demonstrate this, consider step 3 of the algorithm. Since the ergodic probability of state 1 is $O(\epsilon)$, we would proceed with step 3.2. Part (a) of this step tells us to assign $\tilde{\pi}^{(k)}_{j|m}(\epsilon) = \pi^{(k)}_{j|m}(0)$ for each j of interest such that $\pi_{j|m}(0) = O(1)$. State 1 is the only state of interest in this case (it is the state from which the transition that we are counting originates). We proceed with step (b), which requires us to write the equation

$$\tilde{\pi}^{(0)}_{1|m}(\epsilon) = \sum_{i\neq 1}\left(\frac{\lambda_{1i}\tilde{\pi}^{(0)}_{i|m}(\epsilon)}{\sum_{p\neq 1}\lambda_{p1}}\right) \tag{4.152}$$

from which we obtain (by substituting the $O(1)$ probabilities given in $\underline{U}^{(0)}(0)$),

$$\tilde{\pi}^{(0)}_{1|m}(\epsilon) \;=\; \frac{\lambda_{12}}{\lambda_{21}+\lambda_{31}}\pi^{(0)}_{2|m}(0) + \frac{\lambda_{43}}{\lambda_{21}+\lambda_{31}}\pi^{(0)}_{3|m}(0)$$

$$= \frac{\epsilon\lambda_{12}}{\lambda_{21}}\pi^{(0)}_{2|m}(0) + \frac{\epsilon\mu_{42}}{\lambda_{21}}\pi^{(0)}_{3|m}(0), \tag{4.153}$$

where only the leading order terms of the coefficients of the probabilities have been kept since we are calculating only the leading order term, $\tilde{\pi}^{(0)}_{1|m}(\epsilon)$. Comparing this to equation (4.69) we obtain

$$\begin{aligned} b^{(0)}_{12} &= \frac{\epsilon\lambda_{12}}{\lambda_{21}} \\ b^{(0)}_{13} &= \frac{\epsilon\mu_{42}}{\lambda_{21}} \\ \text{and } b^{(0)}_{14} &= 0. \end{aligned} \tag{4.154}$$

Finally, completing part (d) of step 3.2, we use equation (4.70), which applied to this case yields

$$\left[\underline{\tilde{Q}}^{(0)}\right]_j = \begin{cases} \left[\underline{Q}^{(0)}\right]_j + b^{(0)}_{1j}\left[\underline{Q}^{(0)}\right]_1, & j \neq 1 \\ 0, & \mathrm{j} = 1. \end{cases} \tag{4.155}$$

If we write this out for each element of the matrix we obtain

$$\begin{aligned} \left[\underline{\tilde{Q}}^{(0)}\right]_4 = \left[\underline{\tilde{Q}}^{(0)}\right]_1 &= 0 \\ \left[\underline{\tilde{Q}}^{(0)}\right]_2 &= \frac{\epsilon^2\lambda_{12}\mu_{31}}{\lambda_{21}} \\ \left[\underline{\tilde{Q}}^{(0)}\right]_3 &= \frac{\epsilon^2\mu_{42}\mu_{31}}{\lambda_{21}} \end{aligned} \tag{4.156}$$

or

$$\underline{\tilde{Q}}^{(0)} = \left[0 \; \frac{\epsilon^2\lambda_{12}\mu_{31}}{\lambda_{21}} \; \frac{\epsilon^2\mu_{42}\mu_{31}}{\lambda_{21}} \; 0\right]. \tag{4.157}$$

Now step 4 can be completed and we obtain

$$\begin{aligned} \underline{Q}^{(1)} &= \underline{\tilde{Q}}^{(0)}\underline{U}^{(0)} \\ \underline{Q}^{(1)} &= \left[\; \frac{\epsilon^2\lambda_{12}\mu_{31}}{\lambda_{21}} \quad \frac{\epsilon^2\mu_{42}\mu_{31}\lambda_{34}}{\lambda_{21}(\lambda_{43}+\lambda_{34})} \;\right]. \end{aligned} \tag{4.158}$$

Now for the next time scale, aggregate states $1^{(1)}$ and $2^{(1)}$ have O(1) ergodic probabilities so we directly obtain

$$\begin{aligned} \underline{Q}^{(2)} &= \underline{\tilde{Q}}^{(1)}\underline{U}^{(1)}(0) \\ &= \underline{Q}^{(1)}\underline{U}^{(1)}(0) \\ &= \left[\ \epsilon^2 \frac{\mu_{31}\lambda_{12}(\mu_{42}\lambda_{34}+\lambda_{43}\mu_{24})+\mu_{13}\mu_{42}\mu_{31}\lambda_{34}}{\lambda_{21}(\mu_{42}\lambda_{34}+\mu_{24}\lambda_{43})+\mu_{13}\lambda_{21}(\lambda_{43}+\lambda_{34})}\ \right]. \end{aligned} \tag{4.159}$$

We therefore see that we have obtained non-zero values for the expected frequencies. Furthermore, note that for the case of $\epsilon = 0$ the elements of $\underline{Q}^{(1)}$ and $\underline{Q}^{(2)}$ do in fact reduce to 0. However given a sufficiently long time scale these frequencies may become significant.

A proof that the values we obtain are the required expected frequencies now follows.

Proof of Algorithm for Almost Transient States

The proof of the validity of our algorithm for calculating transition frequencies which originate in almost transient states proceeds as follows. First, we generalize the concepts of Lemmas 2 and 3 to obtain an expression for the expected frequency of the transitions involving the transition rate for a particular transition of interest and the ergodic probability of the state from which the transition originates.

Next Lemma 6 shows that the calculations in step 3.2 of the algorithm are equivalent to calculating the leading order terms of the conditional ergodic probability matrices. Lemma 7 then obtains a compact expression for the operations of step 3.2 of our algorithm and hence for the matrices $\underline{\tilde{Q}}^{(k)}$. The results of Lemmas 6 and 7 are combined in Lemma 8 to show that the generation of the matrix $\underline{\tilde{Q}}^{(k)}$ in step 3.2 can be replaced by using the leading order terms of the exact ergodic probability matrix in our calculations.

Finally, using the results of these four Lemmas, the result of Theorem 2 can be extended in Theorem 1 to cases of almost transient states. The proof of the theorem substitutes the results of Lemmas 6 through 8 into the result of Lemma 5. The appropriate initial condition for the state probability vector is then used to obtain the desired result.

Starting with Lemma 5 we calculate the expected frequency of a transition as the product of the ergodic probability of the state of origin and the transition rate. The Lemma is basically a generalization of Lemmas 2 and 3.

Lemma 5

Suppose that we have a Markov chain in which we are interested in transitions from state I to state J, and let $\eta(t)$ be the associated counting process. (Note that our standard notation for counting processes indicates that we should use $\eta_{JI}(t)$ here, but since we are counting only a single transition, we drop the subscripts for simplicity.)

If $O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k})$ then

$$\lim_{\epsilon \to 0} \frac{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right] - \lambda_{JI}\pi_{I|m}^{(0,k)}(\epsilon)}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right]} = 0 \tag{4.160}$$

<u>Proof</u> We prove the result by applying the result of Kielson [16] which was applied to a pair of states (I,J') in equation (4.99). If we apply the result here to a pair of states I and J, we obtain

$$\frac{d}{dt}\Pr\{T(t) = (I,J)\} = -\sum_{\substack{i \neq J \\ i=1}}^{n} \lambda_{iJ} \Pr\{\rho(t) = J\} + \lambda_{JI} \Pr\{\rho(t) = I\} \tag{4.161}$$

where T(t) = (i,j) if the last transition prior to time t was from state i to state j.

Now let us derive a process $\bar{T}(t)$ such that

$$\bar{T}(t) = \begin{cases} 1, & T(t) = (I,J) \\ 0, & T(t) \neq (I,J) \end{cases} . \tag{4.162}$$

Then clearly

$$E\left[\bar{T}(t)\right] = \Pr\{T(t) = (I,J)\} \tag{4.163}$$

and from (4.161)

$$\frac{d}{dt}E\left[\bar{T}(t)\right] = -\sum_{\substack{i \neq J \\ i=1}}^{n} \lambda_{iJ} \Pr\{\rho(t) = J\} + \lambda_{JI} \Pr\{\rho(t) = I\}. \tag{4.164}$$

Note that when $\bar{T}(t)$ changes from 0 to 1, $\eta(t)$ increases by 1; however, when $\bar{T}(t)$ changes from 1 to 0, $\eta(t)$ does not change. Then clearly from (4.164),

$$\frac{d}{dt}E\left[\eta(t)\right] = \lambda_{JI} \Pr\{\rho(t) = I\}. \tag{4.165}$$

Now, if we integrate equation (4.165) from time t to t+Δt and divide by Δt, we obtain

$$E\left[\frac{\eta(t+\Delta t) - \eta(t)}{\Delta t}\right] = \int_{t}^{t+\Delta t} \frac{\lambda_{JI} \Pr\{\rho(\tau) = I\} d\tau}{\Delta t}. \tag{4.166}$$

Now suppose that Δt satisfies $O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k})$. For this range of Δt, we have by definition that

$$\pi^{(0,k)}_{i|m}(\epsilon) = \lim_{\substack{\epsilon \to 0 \\ O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k})}} \left[\Pr(\rho(t+\Delta t) = I \mid \rho(t) \in m^{(k)}\right]. \tag{4.167}$$

Substituting this expression into (4.166) we obtain for $O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k})$,

$$E\left[\frac{\eta(t+\Delta t) - \eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right] = \frac{1}{\Delta t}\int_{t}^{t+\Delta t} \lambda_{JI}\pi^{(0,k)}_{I|m}(\epsilon) d\tau (1 + o(1)). \tag{4.168}$$

Integrating (4.168) and rearranging we obtain

$$\begin{aligned} \lim_{\epsilon \to 0} \frac{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right] - \lambda_{JI}\pi^{(k)}_{I|m}(\epsilon)}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right]} &= \lim_{\epsilon \to 0} o(1) \\ &= 0, \end{aligned} \tag{4.169}$$

and we are done.

In the proof of the next lemma (6) we will commonly refer to the states which have O(1) conditional ergodic probabilities, i.e. states in $i^{(k)}$ for which $\pi_{i|m}^{(k)}(\epsilon)$ is O(1) for some $m^{(k+1)}$. Therefore, as before, we define the set of these states at time scale k to be $M^{(k)}$. In addition, when we refer to an arbitrary element of this set, it will be referred to as state R (i.e. recurrent). Similarly, if we refer to a state which has an O(ϵ) conditional ergodic probability, then we refer to such an arbitrary state as state A (almost transient). If an arbitrary state in the chain, which could have either O(1) or O(ϵ) probability is referenced, it will be denoted by a lower case letter such as i or j.

The result in Lemma 6 shows that the steps of the algorithm in Section 4.5.3 are in fact equivalent to calculating the leading order terms of the conditional probabilities. Specifically, there is a set of coefficients, denoted $b_{ji}^{(k)}$, calculated in the algorithm of Section 4.5.3. These coefficients are used to express the ergodic probabilities of states A, which have O(ϵ)ergodic probabilities, as a linear combination of the O(1) ergodic probabilities of states $R \in M^{(k)}$. We show that this set of coefficients can be used in combination with the matrices $\underline{U}^{(k)}(0)$ to find the leading order terms of the elements of the matrix $\underline{U}^{(k)}(\epsilon)$. The matrix of these elements was defined earlier to be $\underline{\tilde{U}}^{(k)}(\epsilon)$. We note here that like the $\underline{Q}^{(k)}$ matrices, the $b_{ji}^{(k)}$ coefficients and the matrix $\underline{B}^{(k)}$ which is generated from these terms are always assumed to be functions of ϵ. Therefore, as in the case of the $\underline{Q}^{(k)}$ matrices, the argument is dropped for simplicity.

Lemma 6

Suppose that we have a Markov chain for which a time scale decomposition has been performed, yielding the ergodic probability matrices $\underline{U}^{(k)}(0)$. Then, performing step 3.2 of our algorithm is equivalent to calculating the leading order terms of $\underline{U}^{(k)}(\epsilon)$, the exact conditional ergodic probabilities. These probabilities are just the limiting probabilities (as $t \rightarrow \infty$) of the model of the system at the kth time scale. The model at this time scale is described by $\underline{A}^{(k)}(\epsilon)$.

Repeating the calculations of step 3.2 of the algorithm, using the notation of A for almost transient states and R for recurrent states, we obtain:

(a) For each $R^{(k)}$ $(\pi^{(k)}_{R|m}(0) = \mathrm{O}(1))$, set

$$\tilde{\pi}^{(k)}_{R|m}(\epsilon) = \pi^{(k)}_{R|m}(0) \tag{4.170}$$

(b) For each state $A^{(k)}$ of interest, write the expression

$$\tilde{\pi}^{(k)}_{A|m}(\epsilon) = \sum_{\substack{i=1\\ i\neq A}}^{n_k} \left(\frac{\lambda^{(k)}_{Ai}\pi^{(k)}_{i|m}(\epsilon)}{\sum_{\substack{j\neq A\\ j=1}}^{n_k} \lambda^{(k)}_{jA}} \right) \tag{4.171}$$

where we recall that n_k is the number of states in the model of the chain at time scale k.

(c) If the right hand side of the expression in (b) involves other terms $\tilde{\pi}^{(k)}_{A|m}(\epsilon)$ that have not been specified in (4.170), write (4.171) for these as well. (Note that terms $\tilde{\pi}^{(k)}_{A|m}(\epsilon)$ are not specified by (a) because they represent ergodic probabilities of almost transient states)

(d) Repeat step (c) until a closed set of equations is obtained.

By solving these equations we express each $\tilde{\pi}^{(k)}_{A|m}(\epsilon)$ as a linear combination of the probabilities $\pi^{(k)}_{R|m}(0)$. Now if we let the coefficient of $\pi^{(k)}_{R|m}(0)$ in the expression for each almost transient probability $\tilde{\pi}^{(k)}_{A|m}(\epsilon)$ be $b^{(k)}_{AR}$, then recalling that $M^{(k)}$ is the set of states with O(1) ergodic probability, we write

$$\tilde{\pi}^{(k)}_{A|m}(\epsilon) = \sum_{R \in M^{(k)}} b^{(k)}_{AR}(\epsilon)\pi^{(k)}_{R|m}(0). \tag{4.172}$$

Up until this point we have only defined $b^{(k)}_{ji}$ for $i^{(k)} \notin M^{(k)}$ and $j^{(k)} \in M^{(k)}$. It is necessary that we define $b^{(k)}_{ji}$ for all i and j so that we can express (4.172) in matrix form. Therefore we define $b^{(k)}_{ji}$ for each remaining combination of i and j, $1 \leq i, j \leq n^{(k)}$ according to

$$b^{(k)}_{ji} = \begin{cases} 0, & i^{(k)} \notin M^{(k)} \\ 0, & j^{(k)} \in M^{(k)} \text{ and } i \neq j \\ 1, & j^{(k)} \in M^{(k)} \text{ and } i = j. \end{cases} \tag{4.173}$$

A word of explanation is due here. First since we are expressing the ergodic probabilities in terms of O(1) probabilities, $b^{(k)}_{ji} = 0$ for all states $i^{(k)}$ that are almost transient. This is expressed by the first line of (4.173). Second, expressing an O(1) ergodic probability as a combination of O(1) probabilities is trivial, because we just set the probability equal to itself. This is represented by lines 2 and 3 in equation (4.173).

Finally, we may form $\underline{B}^{(k)}$ such that

$$\left[\underline{B}^{(k)}(\epsilon)\right]_{ji} = b^{(k)}_{ji}(\epsilon) \tag{4.174}$$

and obtain

$$\tilde{\underline{U}}^{(k)}(\epsilon) = \underline{B}^{(k)}\underline{U}^{(k)}(0). \tag{4.175}$$

Proof

The approach we will take to this proof is to first show that the equations represented by (4.171) do yield the leading order terms of the probabilities. This will be done by writing equations for the exact ergodic probabilities and then reducing them to the leading order case. We start with the probability mass balance equation at state A under steady state conditions. At the kth time scale, we obtain

$$\pi^{(k)}_{A|m}(\epsilon)\sum_{\substack{j=1\\ j\neq A}}^{n_k}\lambda^{(k)}_{jA} = \sum_{\substack{i=1\\ i\neq A}}^{n_k}\lambda^{(k)}_{Ai}\pi^{(k)}_{i|m}(\epsilon) \tag{4.176}$$

where we reall that n_k is the number of states in the model of the chain at time scale k. This can be solved to obtain

$$\pi^{(k)}_{A|m}(\epsilon) = \sum_{\substack{i=1\\ i\neq A}}^{n_k}\frac{\pi^{(k)}_{i|m}(\epsilon)\lambda^{(k)}_{Ai}}{\sum_{\substack{j=1\\ j\neq A}}^{n_k}\lambda^{(k)}_{jA}} \tag{4.177}$$

If there are terms on the right hand side of (4.177) such that $\pi^{(k)}_{i|m}(\epsilon)\neq O(1)$ we write (4.177) for those terms as well. Continuing until a closed form is obtained, we solve to obtain

$$\pi^{(k)}_{I|m}(\epsilon) = \sum_{R\in M^{(k)}} b^{(k)}_{IR}\pi^{(k)}_{R|m}(\epsilon), \tag{4.178}$$

where $b^{(k)}_{IR}$ is the coefficient obtained from the solution of the equations obtained from (4.177). Note that these steps are identical to those defined by equations (4.68) and (4.69) in step 3.2 of our algorithm except we are writing the equations here involving the exact probabilities $\pi^{(k)}_{i|m}(\epsilon)$. Now from our definition, $\pi^{(k)}_{R|m}(\epsilon) =$ O(1), and hence we know that

$$\pi^{(k)}_{R|m}(\epsilon) = \pi^{(k)}_{R|m}(0)(1+O(\epsilon)). \tag{4.179}$$

Therefore, substituting into (4.178),

$$\pi^{(k)}_{A|m}(\epsilon) = \sum_{R\in M^{(k)}} b^{(k)}_{AR}\pi^{(k)}_{R|m}(0)(1+O(\epsilon)) \tag{4.180}$$

which may be rearranged to yield

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{\pi^{(k)}_{A|m}(\epsilon)-\sum_{R\in M^{(k)}} b^{(k)}_{AR}\pi^{(k)}_{R|m}(0)}{\pi^{(k)}_{A|m}(\epsilon)} &= \lim_{\epsilon\to 0}\frac{O(\epsilon)}{1+O(\epsilon)}\\ &= 0. \end{aligned}\tag{4.181}$$

Therefore the leading order term of $\pi^{(k)}_{A|m}(\epsilon)$ is given by

$$\tilde{\pi}^{(k)}_{A|m}(\epsilon) = \sum_{R \in M^{(k)}} b^{(k)}_{AR} \pi^{(k)}_{R|m}(0). \tag{4.182}$$

Now if we wish to express the sum with all n_k states in the summand, including those states which are almost transient, we can assign coefficients corresponding to transient states values of 0. Therefore we let

$$b^{(k)}_{Ai} = 0, \text{ for i} \notin \mathrm{M}^{(\mathrm{k})} \tag{4.183}$$

to obtain all states in the summand and therefore

$$\tilde{\pi}^{(k)}_{A|m}(\epsilon) = b^{(k)}_{Ai} \pi^{(k)}_{i|m}(0). \tag{4.184}$$

The reason we want the probabilities of every state in the summand, is that we will eventually express the relationship in matrix form. For this reason we also wish to obtain expressions in the form of (4.184) for states $R^{(k)} \in M^{(k)}$ on the left side. Now, given the definition of states $R^{(k)}$, $R \in M^{(k)}$, we have that

$$\tilde{\pi}^{(k)}_{R|m}(\epsilon) = \pi^{(k)}_{R|m}(0). \tag{4.185}$$

Therefore, if we wish to express a state $R^{(k)}$ in the form of (4.184) we obtain a trivial set of coefficients,

$$b^{(k)}_{Ri} = \begin{cases} 1, & i^{(k)} = R^{(k)} \\ 0, & \text{otherwise} \end{cases} \tag{4.186}$$

and obtain for any state $i^{(k)}$, regardless of its ergodic probability,

$$\tilde{\pi}^{(k)}_{j|m}(\epsilon) = \sum_{i=1}^{n_k} b^{(k)}_{ji} \pi^{(k)}_{i|m}(0). \tag{4.187}$$

Now we may note from Rohlicek [23] and algorithm 2.1, that the ergodic probability of state $i^{(k)}$ given class $m^{(k+1)}$ is just $\left[\underline{U}^{(k)}(\epsilon)\right]_{im}$. Therefore

$$\begin{aligned} \left[\underline{U}^{(k)}(\epsilon)\right]_{im} &= u^{(k)}_{im}(\epsilon) \\ &= \pi^{(k)}_{i|m}(\epsilon). \end{aligned} \tag{4.188}$$

Similarly

$$\begin{aligned} \tilde{u}^{(k)}_{im}(\epsilon) &= \tilde{\pi}^{(k)}_{i|m}(\epsilon) \\ u^{(k)}_{im}(0) &= \pi^{(k)}_{i|m}(0). \end{aligned} \tag{4.189}$$

Substituting these expressions into (4.184), we obtain

$$\tilde{u}^{(k)}_{j|m}(\epsilon) = \sum_{i=1}^{n_k} b^{(k)}_{ji} u^{(k)}_{i|m}(0). \tag{4.190}$$

Finally, define a matrix $\underline{B}^{(k)}$ such that

$$\left[\underline{B}^{(k)}\right]_{Ai} = b_{Ai}. \tag{4.191}$$

We can then write (4.190) in matrix form to obtain

$$\begin{bmatrix} \tilde{u}^{(k)}_{1|1}(\epsilon) & \cdots & \tilde{u}^{(k)}_{1|n_{k+1}}(\epsilon) \\ \vdots & \ddots & \vdots \\ \tilde{u}_{n_k|1}(\epsilon) & \cdots & \tilde{u}^{(k)}_{n_k|n_{k+1}}(\epsilon) \end{bmatrix} = \begin{bmatrix} b^{(k)}_{11} & \cdots & b^{(k)}_{1n_k} \\ \vdots & \ddots & \vdots \\ b^{(k)}_{n_k1} & \cdots & b^{(k)}_{n_kn_k} \end{bmatrix} \begin{bmatrix} u^{(k)}_{1|1}(0) & \cdots & u^{(k)}_{1|n_{k+1}}(0) \\ \vdots & \ddots & \vdots \\ u^{(k)}_{n_k|1}(0) & \cdots & u^{(k)}_{n_k|n_{k+1}}(0) \end{bmatrix} \tag{4.192}$$

or simplify to obtain

$$\underline{\tilde{U}}^{(k)}(\epsilon) = \underline{B}^{(k)}\underline{U}^{(k)}(0). \tag{4.193}$$

Our next step is to show in Lemma 7, that this same matrix $\underline{B}^{(k)}$ can be used to relate the matrices $\underline{Q}^{(k)}$ and $\underline{\tilde{Q}}^{(k)}$.

Lemma 7

Suppose we have performed a time scale decomposition on a Markov chain and form the matrix $\underline{\tilde{Q}}^{(k)}$ from the matrix $\underline{Q}^{(k)}$ according to equation (4.70) of our algorithm. Then this is equivalent to postmultiplying by $\underline{B}^{(k)}$ as defined in Lemma 6. Specifically

$$\underline{\tilde{Q}}^{(k)} = \underline{Q}^{(k)}\underline{B}^{(k)} \tag{4.194}$$

Proof In Lemma 6, we saw that equation (4.187) held. That is,

$$\tilde{\pi}^{(k)}_{j|m}(\epsilon) = \sum_{i1}^{n_k} b^{(k)}_{ji}\pi^{(k)}_{i|m}(0). \tag{4.195}$$

However, from our algorithm (specifically equation (4.70)) we define the matrix $\underline{\tilde{Q}}^{(k)}$ using

$$\begin{array}{rcll} \left[\underline{\tilde{Q}}^{(k)}\right]_R & = & \left[\underline{Q}^{(k)}\right]_R + \sum_{A^{(k)}\notin M^{(k)}} b^{(k)}_{AR}\left[\underline{Q}^{(k)}\right]_A, & R \in M^{(k)} \\ \left[\underline{\tilde{Q}}^{(k)}\right]_A & = & 0, & A \notin M^{(k)} \end{array} \tag{4.196}$$

Again we would like a sum that includes all states. Therefore we proceed by noting from the definition of $b^{(k)}_{ji}$ in Lemma 6 that if $j^{(k)} \in M^{(k)}$,

$$b_{jR} = \begin{cases} 1 & j^{(k)} = R^{(k)} \\ 0 & j^{(k)} \neq R^{(k)}, \end{cases} \tag{4.197}$$

which implies that

$$\sum_{j^{(k)}\in M^{(k)}} b^{(k)}_{jR}\left[\underline{Q}^{(k)}\right]_j = \left[\underline{Q}^{(k)}\right]_R. \tag{4.198}$$

This expression can then be substituted into (4.196) to yield

$$\begin{aligned} \left[\underline{\tilde{Q}}^{(k)}\right]_R &= \sum_{j^{(k)}\in M^{(k)}} b^{(k)}_{jR}\left[\underline{Q}^{(k)}\right]_j + \sum_{j^{(k)}\notin M^{(k)}} b^{(k)}_{jR}\left[\underline{Q}^{(k)}\right]_j \\ &= \sum_{j=1}^{n_k} b^{(k)}_{jR}\left[\underline{Q}^{(k)}\right]_j. \end{aligned} \tag{4.199}$$

However, since $\left[\underline{\tilde{Q}}^{(k)}\right]_A = 0$ and $b^{(k)}_{iA} = 0$ for all $j^{(k)}$, equation (4.199) holds for all states. This equation can be written in matrix form to obtain

$$\underline{\tilde{Q}}^{(k)} = \underline{Q}^{(k)}\underline{B}^{(k)}. \tag{4.200}$$

Using this result and the fact that $\underline{\tilde{U}}^{(k)}(\epsilon) = \underline{B}^{(k)}\underline{U}^{(k)}(0)$, from Lemma 6, we wish to show that we may form the matrices $\underline{Q}^{(k)}$ using either $\underline{\tilde{Q}}^{(k)}$ or the $\underline{\tilde{U}}^{(k)}(\epsilon)$ matrices and obtain equivalent results.

Lemma 8

Suppose we have a Markov chain for which a time scale decomposition has been performed yielding the ergodic probability matrices $\underline{U}^{(k)}(0)$. Furthermore let the matrices $\underline{Q}^{(k)}(\epsilon)$ be calculated using the algorithm of Section 4.5.3, while

$$\hat{\underline{Q}}^{(k)} = \underline{Q}^{(0)}\tilde{\underline{U}}^{(0)}(\epsilon)\tilde{\underline{U}}^{(1)}(\epsilon)\cdots\tilde{\underline{U}}^{(k-1)}(\epsilon). \tag{4.201}$$

Then we have that

$$\underline{Q}^{(k)} = \hat{\underline{Q}}^{(k)} \tag{4.202}$$

Proof:

From Lemma 7 we have that

$$\tilde{\underline{Q}}^{(k)} = \underline{Q}^{(k)}\underline{B}^{(k)} \tag{4.203}$$

and from the algorithm (equation 4.76) we have that

$$\underline{Q}^{(k)} = \tilde{\underline{Q}}^{(k-1)}\underline{U}^{(k-1)}(0). \tag{4.204}$$

If we substitute equation (4.203) into equation (4.204), and then re-substitute the results corresponding to k-1, k-2, ...,2, 1, we obtain

$$\begin{aligned}
\underline{Q}^{(k)} &= \underline{Q}^{(k-1)}\underline{B}^{(k-1)}\underline{U}^{(k-1)}(0) \\
&= \underline{Q}^{k-2}\underline{B}^{(k-2)}\underline{U}^{(k-2)}(0)\underline{B}^{(k-1)}\underline{U}^{(k-1)}(0) \\
&\vdots \\
&= \underline{Q}^{(0)}\ \left[\underline{B}^{(0)}(0)\underline{U}^{(0)}(0)\right]\ \left[\underline{B}^{(1)}\ \underline{U}^{(1)}(0)\right]\cdots\left[\underline{B}^{(k-1)}\ \underline{U}^{(k-1)}\right] \\
&= \underline{Q}^{(0)}\tilde{\underline{U}}^{(0)}(\epsilon)\tilde{\underline{U}}^{(1)}(\epsilon)\cdots\tilde{\underline{U}}^{(k-1)}(\epsilon) \\
&= \hat{\underline{Q}}^{(k)}
\end{aligned} \tag{4.205}$$

where we have used Lemma 6 to obtain the second last line of (4.205).

We are now ready to prove Theorem 1.

Theorem 1:

Suppose that we have a Markov chain such that we are interested in counting a set of transitions and there is a count process $\eta(t)$ equal to the number of transitions up to time t. Furthermore, let us form the matrices $\underline{Q}^{(k)}$ using the algorithm of Section 4.5.3. If $O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k})$ then

$$\lim_{\epsilon \to 0} \frac{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right] - \left[\underline{Q}^{(k)}\right]_{m}}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right]} = 0 \tag{4.206}$$

Proof:

Start with the case where there is a single transition to count and let this transition be from a state I to a state J. If state I is such that $\pi_I(\epsilon)$ is O(1), then we are done because Theorem 2 applies. If however we have $\pi_I(\epsilon) = o(1)$ then we proceed as follows.

First we recognize that the leading order term of a conditional ergodic probability $(\tilde{\pi}_{I|m}^{(0,k)}(\epsilon))$ can be calculated using the leading order terms of the ergodic probability matrices $(\tilde{\underline{U}}^{(k)}(\epsilon))$. To see this we note that for $O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k})$,

$$\begin{aligned} \lim_{\epsilon \to 0} \underline{x}(t+\Delta t) &= \underline{U}^{(0)}(\epsilon) \ldots \underline{U}^{(k-1)}(\epsilon) \underline{V}^{(k-1)}(\epsilon) \ldots \underline{V}^{(0)}(\epsilon) \underline{x}(t) \\ &= \left[\tilde{\underline{U}}^{(0)}(\epsilon)(1+O(\epsilon))\right] \ldots \left[\tilde{\underline{U}}^{(k-1)}(\epsilon)(1+O(\epsilon))\right] \ldots \\ &\quad \left[\tilde{\underline{V}}^{(k-1)}(\epsilon)(1+O(\epsilon))\right] \ldots \left[\tilde{\underline{V}}^{(0)}(\epsilon)(1+O(\epsilon))\right] \underline{x}(t). \end{aligned} \tag{4.207}$$

Furthermore, if we let

$$\underline{x}^{(k)}(t) = \tilde{\underline{V}}^{(k-1)}(\epsilon) \ldots \tilde{\underline{V}}^{(0)}(\epsilon) \underline{x}(t) \tag{4.208}$$

then we obtain

$$\lim_{\substack{\Delta t \to \infty \\ O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k})}} \underline{x}(t+\Delta t) = \tilde{\underline{U}}^{(0)}(\epsilon) \ldots \tilde{\underline{U}}^{(k-1)}(\epsilon) \underline{x}^{(k)}(t)(1+O(\epsilon)). \tag{4.209}$$

Given (4.209), recall that the row vector $\left[\underline{Q}^{(0)}\right]_i$ is defined as (for a single transition from state I to state J)

$$\left[\underline{Q}^{(0)}\right]_i = \begin{cases} \lambda_{JI}, & i = I \\ 0, & \text{otherwise}, \end{cases} \tag{4.210}$$

according to equation (4.62) in step 2 of our algorithm. If we premultiply both sides of equation (4.209) by $\underline{Q}^{(0)}$, the result is therefore

$$\begin{aligned} \lambda_{JI}\pi^{(0,k)}_{I|m}(\epsilon) &= \underline{Q}^{(0)}\underline{\tilde{U}}^{(0)}(\epsilon)\ldots\underline{\tilde{U}}^{(k-1)}(\epsilon)\underline{x}^{(k)}(t)(1+O(\epsilon)) \\ &= \underline{\bar{Q}}^{(k)}\underline{x}^{(k)}(t)(1+O(\epsilon)) \\ &= \underline{Q}^{(k)}\underline{x}^{(k)}(t)(1+O(\epsilon)), \end{aligned} \tag{4.211}$$

We may further simplify this expression by using the the initial condition that $\rho(t) \in m^{(k)}$. We therefore obtain the row vector $\underline{x}^{(k)}(t)$ as

$$\left[\underline{x}^{(k)}(t)\right]_i = \begin{cases} 1 & i = m \\ 0 & \text{otherwise} \end{cases} \tag{4.212}$$

Now we substitute this initial condition into equation (4.211) to yield

$$\lambda_{JI}\pi^{(0,k)}_{I|m}(\epsilon) = \left[\underline{Q}^{(k)}\right]_m (1+O(\epsilon)). \tag{4.213}$$

We can now rearrange this equation and take the limit as $\epsilon \to 0$ to obtain

$$\lim_{\epsilon\to 0} \frac{\lambda_{JI}\pi^{(0,k)}_{I|m}(\epsilon) - \left[\underline{Q}^{(k)}\right]_m}{\lambda_{JI}\pi^{(0,k)}_{I|m}(\epsilon)} = 0. \tag{4.214}$$

Now we have the result we require beause we can substitute (4.214) into (4.160) to obtain

$$\lim_{\epsilon\to 0} \frac{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right] - \left[\underline{Q}^{(k)}\right]_m}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right]} = 0 \tag{4.215}$$

and therefore we are finished if there is one transition of interest.

To obtain the result for multiple transitions of interest we apply the same arguments that were used in the corollary to Theorem 2. Specifically, suppose we have the conditions of Theorem 1 ($O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k})$) and are counting N individual transitions, represented by counting processes $\eta_1(t)$, $\eta_2(t)$,...,$\eta_N(t)$. Define $\underline{Q}_i^{(k)}$ for each count process as we did in the statement of Theorem 1 for one count process, and then if we are counting all of the transitions collectively, we may obtain the matrix $\underline{Q}^{(k)}$ by calculating

$$\underline{Q}^{(k)} = \sum_{i=1}^{N} \underline{Q}_i^{(k)}. \tag{4.216}$$

Proof If $\eta(t)$ is the total number of transitions that occur, from the mutual exclusivity of transitions, then

$$\eta(t) = \sum_{i=1}^{N} \eta_i(t). \tag{4.217}$$

Hence we may show that

$$\begin{aligned}
\lim_{\epsilon\to 0}\left[\frac{\underline{Q}^{(k)}\underline{x}^{(k)}(t) - E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]}\right] &= \lim_{\epsilon\to 0}\left[\frac{\sum_{i=1}^{N}\underline{Q}_i^{(k)}\underline{x}^{(k)}(t) - \sum_{i=1}^{N}E\left[\frac{\eta_i(t+\Delta t)-\eta_i(t)}{\Delta t}\right]}{\sum_{i=1}^{N}E\left[\frac{\eta_i(t+\Delta t)-\eta_i(t)}{\Delta t}\right]}\right] \\
&\leq \lim_{\epsilon\to 0}\sum_{i=1}^{p}\left[\frac{\underline{Q}_i^{(k)}\underline{x}^{(k)}(t) - E\left[\frac{\eta_i(t+\Delta t)-\eta_i(t)}{\Delta t}\right]}{E\left[\frac{\eta_i(t+\Delta t)-\eta_i(t)}{\Delta t}\right]}\right] \\
&= 0. \qquad (4.218)
\end{aligned}$$

and the Theorem is proved for the case of multiple transitions of interest.

Now that we have Theorem 1, we need only demonstrate that $\underline{Q}^{(k)}$ in that Theorem is generated by the algorithm of Section 4.5.3. Immediately following Theorem 1, we showed that steps 1 and 2 of the algorithm provided the desired initial results. Specifically, the time scale decomposition is first performed to yield the matrices $\underline{A}^{(k)}(\epsilon)$, $\underline{\tilde{V}}^{(k)}(\epsilon)$ and $\underline{U}^{(k)}(0)$. and then Step 2 is used to generate $\underline{Q}^{(0)}$.

Now for step 3, we must consider almost transient states. However, from Lemmas 6 and 7, the calculations in step 3 yield a set of coefficients, $b_{ji}^{(k)}$ such that

$$\left[\underline{B}^{(k)}\right]_{ji} = b_{ji}^{(k)},$$

$$\tilde{\underline{Q}}^{(k)} = \underline{Q}^{(k)}\underline{B}^{(k)}$$

and

$$\tilde{\underline{U}}^{(k)}(\epsilon) = \underline{B}^{(k)}\underline{U}^{(k)}(0).$$

We then found in Lemma 8 that step 4, which required the calculation of $\underline{Q}^{(k+1)} = \tilde{\underline{Q}}^{(k)}\underline{U}^{(k)}(0)$, is equivalent to using the matrix containing the leading order terms of the conditional ergodic probabilities $\tilde{\underline{U}}^{(k)}(\epsilon)$ in the calculations. The result of Theorem 1 then follows for $\underline{Q}^{(k)}$, by applying the initial condition $\rho(t) \in m^{(k)}$. Hence $\underline{Q}^{(k)}$ as calculated by the algorithm satisfies the main result of Theorem 1, or specifically

$$\lim_{\epsilon\to 0} \frac{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right] - \left[\underline{Q}^{(k)}\right]_m}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) \in m^{(k)}\right]} = 0.$$

4.5.5 State Splitting Interpretation

The method of calculating the frequency of transitions originating in almost transient states was discussed and proved in the previous section. The method has a nice analytical interpretation, being equivalent to the calculation of leading order probabilities. However the intuition behind the process requires further discussion.

To provide this intuition, we consider the example that was discussed in the previous section and shown in transition diagram form in Figure 4.18. When the required analysis was performed, the transition frequency matrices were formed to obtain

$$\begin{aligned}
\underline{Q}^{(0)} &= \begin{bmatrix} \mu_{31} & 0 & 0 & 0 \end{bmatrix} && (4.219)\\
\underline{Q}^{(1)} &= \begin{bmatrix} \frac{\epsilon^2\lambda_{12}\mu_{31}}{\lambda_{21}} & \frac{\epsilon^2\mu_{42}\mu_{31}\lambda_{34}}{\lambda_{21}(\lambda_{43}+\lambda_{34})} \end{bmatrix} && (4.220)\\
\underline{Q}^{(2)} &= \begin{bmatrix} \epsilon^2\frac{\mu_{31}\lambda_{12}(\mu_{42}\lambda_{34}+\lambda_{43}\mu_{24})+\mu_{13}\mu_{42}\mu_{31}\lambda_{34}}{\lambda_{21}(\mu_{42}\lambda_{34}+\mu_{24}\lambda_{43})+\mu_{13}\lambda_{21}(\lambda_{43}+\lambda_{34})} \end{bmatrix}. && (4.221)
\end{aligned}$$

However, in the process of finding $\underline{Q}^{(1)}$, another matrix, $\underline{\tilde{Q}}^{(0)}$, was obtained which was given by

$$\underline{\tilde{Q}}^{(0)} = \begin{bmatrix} 0 & \frac{\epsilon^2\mu_{31}\lambda_{12}}{\lambda_{21}} & \frac{\epsilon^2\mu_{31}\mu_{42}}{\lambda_{21}} & 0 \end{bmatrix}. \tag{4.222}$$

Now if we recall the definition of $\underline{Q}^{(0)}$ we note that its ith element is equal to the transition rate for transitions that originate in state i. Therefore, by comparing $\underline{\tilde{Q}}^{(0)}$ to $\underline{Q}^{(0)}$, the operations we have performed appear to have replaced the transition from state 1 by transitions originating in state 2 and 3.

The intuition behind this can be easily explained, but first we introduce a concept known as state splitting that is thoroughly described in Rohlicek [23]. Suppose we have three states in a Markov chain: H, I and J. Furthermore assume that there is a transition from state H to state I and a transition from state I to state J. This situation is illustrated in Figure 4.19. Now suppose that we are interested in a transition *sequence* from state H to state I followed by a transition to state J. Then we may generate a new chain for which state I is split into state $\tilde{I}$ and $\bar{I}$ such that

1) The evolution of the probabilities for states i $\neq$ I is identical to the original chain.

2) If $\rho'(t)$ denotes the derived Markov chain, then

$$\Pr(\rho(t) = I) = \Pr(\rho'(t) = \tilde{I}) + \Pr(\rho'(t) = \bar{I}) \tag{4.223}$$

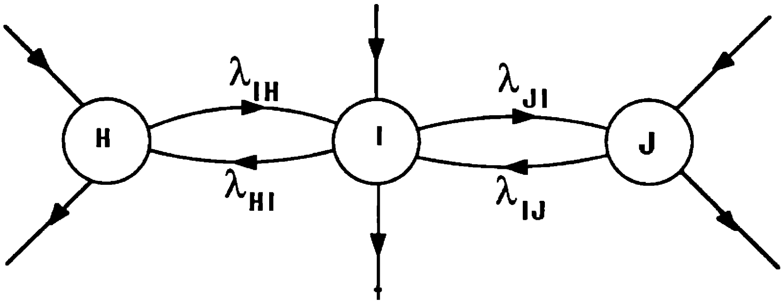


Figure 4.19: Example to Demonstrate State Splitting

3) The only transition originating in state $\tilde{I}$ terminates in state J.

This situation is depicted in Figure (4.20). From [23], (equations (2.85) and (2.86)) we can calculate $\lambda_{\tilde{I}H}$ as the transition rate from state H to state I multiplied by the probability that the next transition leaving state I enters state J. This yields

$$\lambda_{\tilde{I}H} = \lambda_{IH} \frac{\lambda_{JI}}{\sum_{\substack{i=1 \\ i \neq i}}^{n} \lambda_{iI}}. \tag{4.224}$$

In addition, the transition rate $\lambda_{J\tilde{I}}$ is equal to the unconditional transition rate leaving state I or

$$\lambda_{J\tilde{I}} = \sum_{i=1}^{n} \lambda_{iI}. \tag{4.225}$$

Now we apply this to the chain in our example. Suppose that we have the system of Figure 4.18 and that we are still interested in counting the transitions from state 1 to state 3. Furthermore, let us split state 1 into states which correspond to unique transition sequences through the state. For state 1 in this chain there are exactly four such sequences. There are transition sequences from states 2 and 3 which

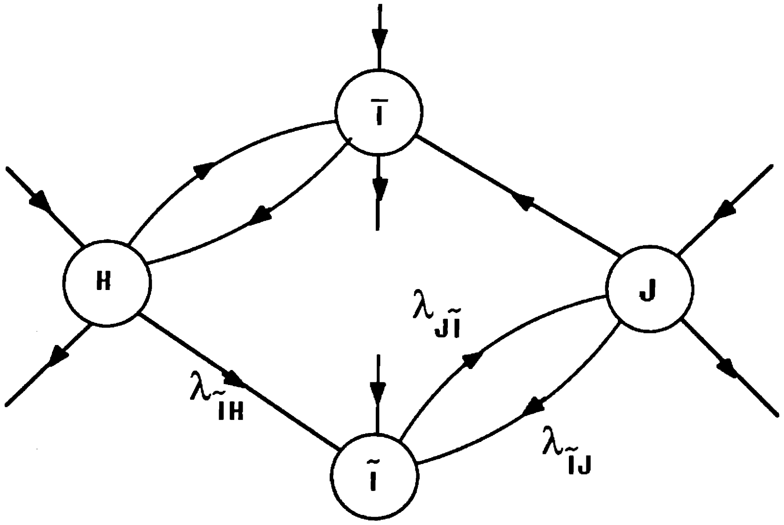


Figure 4.20: New chain with split state

return to the same state as well as sequences from state 2 to state 1 to state 3 or in reverse order. If we use the state splitting techniques described above, we obtain a new chain as illustrated in Figure 4.21.

Our first observation regarding this new chain is that there is a unique path from both state 2 and state 3 which terminates first in a split portion of state 1 and then results in a subsequent transition to state 3. For example a transition from state 2 to state 1B must be followed by a transition from state 1B to state 3.

Our second observation is that the rates for these transitions are equal (to leading order) to the correponding elements in the matrix $\underline{\tilde{Q}}^{(0)}$. (i.e. the transition rate from state 2 to state 1B equals the second element of $\underline{\tilde{Q}}^{(0)}$). In fact, if we recall that the higher order terms in $\underline{\tilde{Q}}^{(0)}$ were already dropped, the agreement is exact. Therefore we see that the formation of $\underline{\tilde{Q}}^{(0)}$ is equivalent to calculating the frequencies of a set of transitions in the new chain (which we constructed via state splitting) that originate in states whose ergodic probability is O(1).

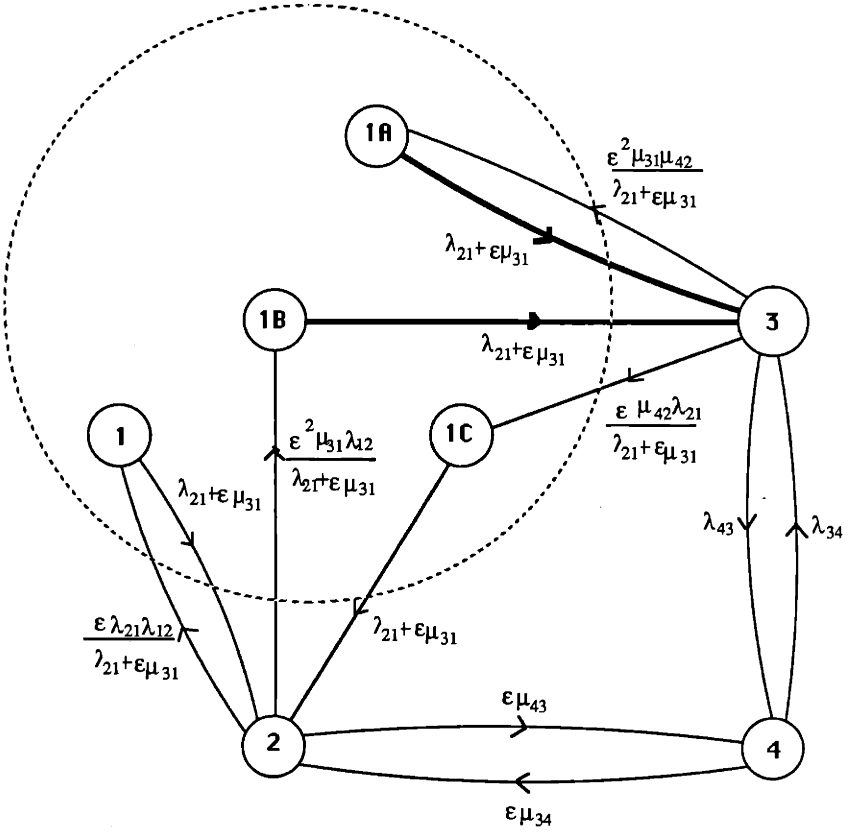

1A
$\frac{\varepsilon^2 \mu_{31}\mu_{42}}{\lambda_{21}+\varepsilon\mu_{31}}$
$\lambda_{21}+\varepsilon\mu_{31}$
1B
3
$\lambda_{21}+\varepsilon\mu_{31}$
1
$\frac{\varepsilon^2 \mu_{31}\lambda_{12}}{\lambda_{21}+\varepsilon\mu_{31}}$
1C
$\frac{\varepsilon\ \mu_{42}\lambda_{21}}{\lambda_{21}+\varepsilon\mu_{31}}$
$\lambda_{21}+\varepsilon\mu_{31}$
λ_{43}
λ_{34}
$\frac{\varepsilon\lambda_{21}\lambda_{12}}{\lambda_{21}+\varepsilon\mu_{31}}$
$\lambda_{21}+\varepsilon\mu_{31}$
$\varepsilon\mu_{43}$
2
4
$\varepsilon\mu_{34}$

Figure 4.21: State Splitting for Almost Transient States

It seems reasonable to suspect at this point that the state splitting approach is an alternative to our approach for handling transitions originating in almost transient states. However, even for this simple four state chain, we required an expansion of the state space to seven states. Since our algorithm does not require such an expansion, we will retain it for calculation purposes. Further justification of this apparent substitution of transitions for calculation purposes is provided in the appendix.

4.5.6 Transition Frequencies over Long Time Intervals

This section deals with the fact that over very long time intervals, a large number of transitions occur and therefore the actual transition frequency observed for a sample path will equal the expected frequency with probability 1. This result is stated precisely by Theorem 3.

Theorem 3

Suppose we have an n-state continuous time Markov chain and we are interested in the frequency of transitions from state I to state J. We then form the transition frequency matrix $\underline{Q}^{(k)}(\epsilon)$ for each time scale using the the transition frequency algorithm. Let the time interval Δt satisfy $O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k})$. Then, if for some aggregate state $m^{(k)}$ we have

$$\left[\underline{Q}^{(k)}\right]_m \succeq O(\epsilon^{k-1}) \tag{4.226}$$

or equivalently

$$\left[\underline{Q}^{(k)}\right]_m \Delta t \succ 1 \tag{4.227}$$

and

$$\rho(t) = m^{(k)}, \tag{4.228}$$

then

$$\lim_{\epsilon \to 0} \frac{\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right] - \left[\underline{Q}^{(k)}\right]_m}{\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]} \ w.p.\ 1. \tag{4.229}$$

i.e. the expected frequency will equal the sample path frequency as $\epsilon \to 0$.

Proof:

Proceed by scaling time using $\tau = t\left[\underline{Q}^{(k)}\right]_m$, which yields

$$\Delta\tau \succ O(1) \tag{4.230}$$

and if the transition frequency matrix at the new time scale is denoted $\underline{\bar{Q}}^{(k)}$, then

$$\left[\underline{\bar{Q}}^{(k)}\right]_m = 1. \tag{4.231}$$

Following the limiting arguments of Lemma 3, let η(t) be the number of transitions of interest and $\bar{X}$ the expected time between transitions. We then obtain

$$\lim_{\epsilon \to 0} \frac{\eta(\tau + \Delta\tau) - \eta(\tau)}{\Delta\tau} = \lim_{\epsilon \to 0} \frac{1}{\bar{X}},\ w.p.\ 1. \tag{4.232}$$

which was proved originally as a consequence of the law of large numbers. Specifically, the sample mean of the time between transitions from state I to state J equals the expected value with probability 1. However, by definition

$$\bar{X} = E\left[\frac{\eta(\tau + \Delta\tau) - \eta(\tau)}{\Delta\tau}\right]^{-1}. \tag{4.233}$$

We may also write an equivalent expression conditioned on the initial state being $m^{(k)}$ to obtain

$$\bar{X}_m^{-1} = E\left[\frac{\eta(\tau + \Delta\tau) - \eta(\tau)}{\Delta\tau} \mid \rho(\tau) \in m^{(k)}\right], \tag{4.234}$$

where $\bar{X}_m$ is the expected time between transitions from state I to state J conditioned on the initial state being in the aggregate state $m^{(k)}$. Then, substituting this expression into Theorem 1, we obtain

$$\lim_{\epsilon \to 0} \frac{\left[\underline{Q}^{(k)}\right]_m - \bar{X}_m^{-1}}{\bar{X}_m^{-1}} = 0. \tag{4.235}$$

Therefore, rescaling to the original time variable, and substituting (4.233),

$$\lim_{\epsilon \to 0} \frac{\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right] - \left[\underline{Q}^{(k)}\right]_m}{\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]} = 0 \tag{4.236}$$

as required.

Examples

We now apply the concepts of this section to two examples.

Example 1

Consider our first example of Section 4.5.2, where we calculated :

$$\underline{Q}^{(0)} = [3\ 0\ 4\ 0]$$

$$\underline{Q}^{(1)} = \left[\frac{6}{5}\ \frac{4}{5}\right]$$

and

$$\underline{Q}^{(2)} = \left[\frac{16}{17}\right]$$

To determine whether or not the results of this section apply to a given event at a given time scale we proceed by forming the product of the expected frequency and the magnitude of Δt at the particular time scale of interest. If the result satisfies (4.227), i.e. it is much larger than 1, then the result of Theorem 3 applies and the sample frequencies will equal the expected frequencies. For our first example, consider the shortest time scale (time scale 0). We are observing the system over time intervals [t,t+Δt] where $\Delta t = o(1)$. The elements of the transition frequency matrix that are of interest are obviously those which are non-zero. From $\underline{Q}^{(0)}$ we see that the non-zero elements are O(1). Therefore the product of the non-zero event frequencies and Δt is o(1) which is small relative to 1 as $\epsilon \to 0$. Hence the result of Theorem 3 does not apply.

Now consider the next time scale, time scale (1). In this case we are considering time intervals such that $\Delta t \succ$ O(1) while the non-zero elements of $\underline{Q}^{(1)}$ are again O(1). Therefore the product of the transition frequency and the time interval is much larger than 1, and hence Theorem 1 applies. For time scale (2) we obtain a similar result since the elements of $\underline{Q}^{(2)}$ are O(1) and the time intervals are even longer ($\succ O(\epsilon^{-1})$).

These results make intuitive sense because increasing the length of the interval makes the expected number of transitions larger. Once the expected number of transitions becomes very large, the sample path frequency approaches the expected frequency with high probability.

<u>Example 2:</u>

Consider the second example of Section 4.5.2. We calculated:

$$\underline{Q}^{(0)} = [3\ 0\ 0\ 0]$$

$$\underline{Q}^{(1)} = \left[2\epsilon\ \frac{2}{5}\epsilon\right]$$

and

$$\underline{Q}^{(2)} = \left[\frac{6}{7}\epsilon\right].$$

Once again we must consider the product of the elements of the matrices $\underline{Q}^{(k)}$ with the length of the time interval Δt. In this case we obtain a similar result at the first time scale because the non-zero elements of $\underline{Q}^{(0)}$ are once again O(1). If we consider the second time scale, however we get a different result. For this example, the non-zero elements of $\underline{Q}^{(1)}$ are O(ϵ). If we multiply this by a time interval of $\Delta t = o(\epsilon^{-1})$ the product is small with respect to 1 (specifically o(1)) and therefore Theorem 3 does not apply. The reason for this result is that the transition originates in an almost transient state. Therefore a very large amount of time must pass before the required large number of transitions take place.

If we examine time scale 2 we see that Δt is long enough at this time scale, for Theorem 3 to apply, since $\Delta t \succ O(\epsilon^{-1})$ and the non-zero elements of the transition frequency matrix are $O(\epsilon)$. Therefore the sample frequencies and expected frequencies will be equal at this time scale.

4.6 Summary

In this chapter, we have discussed procedures for performing time scale decompositions on Markov Chains, lumping Markov Chains and calculating expected frequencies for various transitions. The aggregation techniques provide an exact description of the system as $\epsilon \rightarrow 0$ by generating a number of lower order chains, while the lumping techniques reduce the order of the overall model by eliminating some of the unnecessary details. Transition frequency calculations were then presented. These techniques are useful when systems are modeled in which the frequency of certain events is important. For our case, Flexible Manufacturing Systems, we are interested in the frequency at which parts are produced so that we can match our supply with demand. Finally, in some instances, namely when the interval of observation is long enough for many of these events to occur, the expected frequency will match the actual frequency for a sample path with probability one.

4.A Appendix

4.A.1 State Space Partitioning for Large Models

This appendix describes a method for performing the calculations described in the body of this chapter for chains with a large number of states. The method relies on the fact that many of the calculations require the manipulation of matrices which contain a large number of elements which are identically zero. If we recognize that specific elements in the matrices are non-zero, we can significantly reduce the number of operations required to perform our calculations.

Our method takes advantage of the sparsity of the matrices by splitting or partitioning the state space into a number of subspaces. These subspaces are specifically chosen to reduce the dimension of the matrices which must be manipulated. For example, suppose that we had matrices $\underline{A}$ and $\underline{B}$ such that

$$\underline{A} = \begin{bmatrix} \underline{A}_1 & \underline{0} \\ \underline{0} & \underline{A}_2 \end{bmatrix} \tag{4.237}$$

and

$$\underline{B} = \begin{bmatrix} \underline{B}_1 & \underline{0} \\ \underline{0} & \underline{B}_2 \end{bmatrix} \tag{4.238}$$

i.e.$\underline{A}$ and $\underline{B}$ are block diagonal. If $\underline{A}$ and $\underline{B}$ have dimensions such that we can multiply $\underline{A}$ with $\underline{B}$ and $\underline{A}_i$ with $\underline{B}_i$, then we can clearly find $\underline{A}\ \underline{B}$ by multiplying the lower dimension submatrices and obtain

$$\underline{A}\ \underline{B} = \begin{bmatrix} \underline{A}_1\,\underline{B}_1 & \underline{0} \\ \underline{0} & \underline{A}_2\,\underline{B}_2 \end{bmatrix}. \tag{4.239}$$

This type of approach will be needed for systems of large dimension when the calculations described in Section 4.3 are implemented. For example, suppose that we have a model with n states at the shortest time scale and N ergodic classes. Therefore, equation (4.240) requires the multiplication of Nxn, nxn, and nxN matrices.

$$\underline{A}^{(1)}(\epsilon) = \frac{1}{\epsilon}\, \underline{V}^{(0)}(\epsilon)\, \underline{A}^{(0)}(\epsilon)\, \underline{U}^{(0)}(0) \tag{4.240}$$

These multiplications can be simplified if we realize that some of the matrices required in the calculations exhibit a block diagonal structure. We will find that for the models that we work with, there is a natural partitioning of the state space such that each subspace is associated with a diagonal block in the matrix.

We demonstrate this by considering each of the matrices in (4.240). We start with the transition rate matrix of the Markov chain. If the system displays multiple time scale behavior, then we know that we can order the state such that

$$\underline{A}^{(k)}(\epsilon) = \begin{bmatrix} \underline{A}_{11}^{(k)}(\epsilon) & \cdots & \underline{A}_{1N}^{(k)}(\epsilon) & \underline{A}_{1T}^{(k)}(\epsilon) \\ \vdots & \ddots & \vdots & \vdots \\ \underline{A}_{N1}^{(k)}(\epsilon) & \cdots & \underline{A}_{NN}^{(k)}(\epsilon) & \underline{A}_{NT}^{(k)}(\epsilon) \\ \underline{0} & \cdots & \underline{0} & \underline{A}_{TT}^{(k)}(0) \end{bmatrix} \tag{4.241}$$

where there are N ergodic classes for the chain, and each submatrix $\underline{A}_{ij}^{(k)}(\epsilon)$ is associated with transitions from states in ergodic class j to states in ergodic class i. The submatrices $\underline{A}_{iT}^{(k)}(\epsilon)$ contain the transition rates from transient states into the states of ergodic class i. Furthermore, we know from the multiple time scale behavior of the chain that the elements of the matrices $\underline{A}_{ij}^{(k)}(\epsilon), i \neq j$ are all $O(\epsilon)$. However, the calculations given by (4.240) require $\underline{A}^{(0)}(\epsilon)$, so that the structure of the transition rate matrix itself does not exhibit any block diagonal structure that will help us in our calculations.

For $\underline{U}^{(k)}(0)$, the ergodic probability matrix, we know that for ϵ=0, the ergodic probability of a state given a particular ergodic class is non zero only if the state belongs to that ergodic class. Therefore, $\underline{U}^{(k)}(0)$ can be written in the form

$$\underline{U}^{(k)}(0) = \begin{bmatrix} \underline{U}_{11}^{(k)}(0) & \cdots & \underline{0} \\ \vdots & \ddots & \vdots \\ \underline{0} & \cdots & \underline{U}_{NN}^{(k)}(0) \\ \underline{0} & \cdots & \underline{0} \end{bmatrix} \tag{4.242}$$

Similarly, if we examine the class membership matrix we recognize that the non-transient states belong only to a single aggregate class, but the transient states may belong to a number of ergodic classes. Therefore, we know that we can write $\underline{V}^{(k)}(\epsilon)$ in the form

$$\underline{V}^{(k)}(\epsilon) = \begin{bmatrix} \underline{V}_{11}^{(k)}(\epsilon) & \cdots & \underline{0} & \underline{V}_{1T}^{(k)}(\epsilon) \\ \vdots & \ddots & \vdots & \vdots \\ \underline{0} & \cdots & \underline{V}_{NN}^{(k)}(\epsilon) & \underline{V}_{NT}^{(k)}(\epsilon) \end{bmatrix}. \tag{4.243}$$

However, from Rohlicek [23], we only need to keep the O(ϵ) terms for the transient states, so we may use the form of $\tilde{\underline{V}}^{(k)}(\epsilon)$ given in (4.244).

$$\tilde{\underline{V}}^{(k)}(\epsilon) = \begin{bmatrix} \underline{V}_{11}^{(k)}(0) & \cdots & \underline{0} & \underline{V}_{1T}^{(k)}(\epsilon) \\ \vdots & \ddots & \vdots & \vdots \\ \underline{0} & \cdots & \underline{V}_{NN}^{(k)}(0) & \underline{V}_{NT}^{(k)}(\epsilon) \end{bmatrix} \tag{4.244}$$

We note here that we need not partition our state space exactly according to the ergodic classes to obtain the forms for $\tilde{\underline{V}}^{(k)}(\epsilon)$ and $\underline{U}^{(k)}(0)$ that are shown in equations (4.242) and (4.244). In particular, any partitioning of the state space such that each ergodic class lies within a single subspace will suffice. This is clearly true because we may take a number of the diagonal blocks in (4.242) or (4.244) and combine them into a single block and the total matrix will still be block diagonal.

For example, suppose we have a matrix $\underline{B}$ such that

$$\underline{B} = \begin{bmatrix} \underline{B}_1 & \underline{0} & \underline{0} \\ \underline{0} & \underline{B}_2 & \underline{0} \\ \underline{0} & \underline{0} & \underline{B}_3 \end{bmatrix} . \tag{4.245}$$

If we let

$$\underline{B}_4 \equiv \begin{bmatrix} \underline{B}_1 & \underline{0} \\ \underline{0} & \underline{B}_2 \end{bmatrix} , \tag{4.246}$$

then we can write $\underline{B}$ as

$$\underline{B} = \begin{bmatrix} \underline{B}_4 & \underline{0} \\ \underline{0} & \underline{B}_3 \end{bmatrix} . \tag{4.247}$$

To ensure that each ergodic class lies within a single subspace, we need only ensure that all transitions between the subspaces that we choose are at most $O(\epsilon)$. We may also include some of the transient (and almost transient) states in each subspace. The only restriction is that the transitions from these states must enter only the states from a single subspace. This restriction is due to the fact that associating an almost transient state with a subspace such that there are transitions to other subspaces may result in the elimination of behavior represented by sequences of rare events through the transient states. (See [23] and the effect of almost transient states on the long term behavior of a Markov chain). We proceed now, without loss of generality to assume that we partition the state space according to the individual ergodic and transient classes. To determine how we can simplify the calculation given by equation (4.240), we rewrite that equation in terms of the partitioned submatrices as defined in equations (4.241),(4.242), and (4.244), which yields:

$$\begin{aligned} \underline{A}_{ij}^{(k+1)}(\epsilon) &= \frac{1}{\epsilon} \sum_{r=1}^{N} \tilde{V}_{ir}^{(k)}(\epsilon) \left(\sum_{s=1}^{N} A_{rs}{}^{(k)}(\epsilon) U_{sj}{}^{(k)}(0) \;+\; A_{rT}{}^{(k)}(\epsilon)\; U_{Tj}{}^{(k)}(0) \right) \\ &+ \frac{1}{\epsilon} \tilde{V}_{iT}^{(k)}(\epsilon) \left(\sum_{s=1}^{N} A_{Ts}{}^{(k)}(\epsilon)\; U_{sj}{}^{(k)}(0) \;+\; A_{TT}{}^{(k)}(\epsilon)\; U_{Tj}{}^{(k)}(0) \right) \end{aligned} \tag{4.248}$$

Now, from (4.244) and (4.242),

$$\begin{aligned} \underline{U}_{Tj}^{(k)}(0) &= \underline{0} \\ \underline{U}_{ij}^{(k)}(0) &= \underline{0}, i \neq j \\ \underline{\tilde{V}}_{ij}^{(k)}(\epsilon) &= \underline{0}, j \neq i, T. \end{aligned} \tag{4.249}$$

Substituting these equations into (4.248) we obtain

$$\underline{A}_{ij}^{(k+1)}(\epsilon) = \frac{1}{\epsilon}\left[\tilde{V}_{ii}^{(0)}(\epsilon) A_{ij}{}^{(k)}(\epsilon) U_{jj}{}^{(k)}(0) \;+\; \tilde{V}_{iT}^{(k)}(\epsilon) A_{Tj}{}^{(k)}(\epsilon) U_{jj}{}^{(k)}(0)\right]. \tag{4.250}$$

This is the simplified formula required for the calculation described by (4.240). Note that we need only deal with matrices that are as large as the submatrices of $\underline{\tilde{V}}^{(k)}(\epsilon)$, $\underline{U}^{(k)}(\epsilon)$ and $\underline{A}^{(k)}(\epsilon)$ that are associated with individual subspaces. Finally, we note that if the choice of subspaces in our problem is based on transitions which result from the slowest events in the system being modelled (at most $O(\epsilon)$), the partitioning will be compatible with the conditions described above.

4.A.2 Comments on State Splitting Interpretation

In Section 4.5.5, the method for handling almost transient states by our algorithm was interpreted using the concept of state splitting. Specifically, we chose a chain that was analyzed in Section 4.5.4 and examined the matrices

$$\underline{Q}^{(0)} = [\epsilon\mu_{31}\ 0\ 0\ 0]$$

and

$$\underline{\tilde{Q}}^{(0)} = \left[\begin{array}{cccc} 0 & \frac{\epsilon^2\mu_{31}\lambda_{12}}{\lambda_{21}} & \frac{\epsilon^2\mu_{21}\mu_{12}}{\lambda_{21}} & 0 \end{array}\right].$$

We then generated a new chain from the old chain (Figure 4.19) to obtain the chain in Figure 4.20. This new chain had identical probabilistic behavior to the original for states other than state 1. In addition, the sum of the probabilities of the states obtained by splitting state 1 of the original chain is equal to the probability of state 1 in the original.

In the new chain we noted that each transition path through state 1 has a unique sequence through one of the split portions of the state. In addition, the transition rates out of states 2 and 3 which lead to a state 1 to state 3 transition are equal to the second and third elements of $\underline{\tilde{Q}}^{(0)}$ respectively. However, if we count those transitions, the counting process that is obtained, which we may call $\tilde{\eta}(t)$ is not equivalent to the original counting process $\eta(t)$ because they are counting different transitions. However, we may in fact show that the expected frequency of the transitions represented by the two different counting processes are asymptotically equal, or specifically, for appropriate time intervals (defined more precisely below)

$$\lim_{\epsilon\to 0}\frac{E\left[\frac{\tilde{\eta}(t+\Delta t)-\bar{\eta}(t)}{\Delta t}\right]-E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]}{E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right]}=0. \tag{4.251}$$

We start by considering a somewhat simpler example where we have 3 states H,I, and J as shown in Figure 4.22. From our earlier notation, we have that the transitions from state H to state I are counted by $\eta_{IH}(t)$ and those from state I to state J are counted by $\eta_{JI}(t)$. We wish to show for this case that if $\rho(t) \neq I$, and $\Delta t \succ O(1)$ then

$$\lim_{\epsilon\to 0}\frac{E\left[\frac{\eta_{JI}(t+\Delta t)-\eta_{JI}(t)}{\Delta t}\right]-E\left[\frac{\eta_{IH}(t+\Delta t)-\eta_{IH}(t)}{\Delta t}\right]}{E\left[\frac{\eta_{JI}(t+\Delta t)-\eta_{JI}(t)}{\Delta t}\right]}=0. \tag{4.252}$$

Now, for the Markov chain described above let

$$\begin{aligned}\tilde{\varsigma}(\Delta t) &= \frac{\eta_{IH}(t+\Delta t)-\eta_{IH}(t)}{\Delta t}\\ \varsigma(\Delta t) &= \frac{\eta_{JI}(t+\Delta t)-\eta_{JI}(t)}{\Delta t}\end{aligned} \tag{4.253}$$

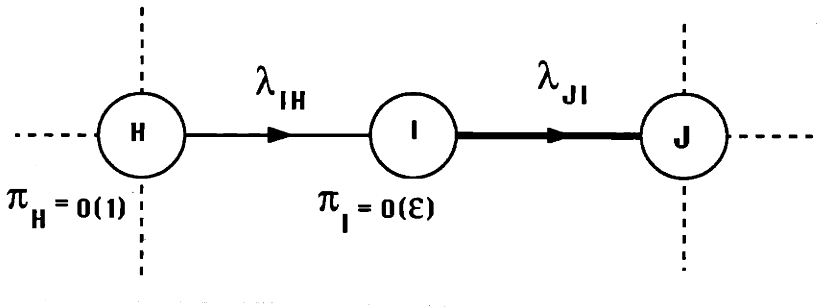


Figure 4.22: Transition Sequence Example

and $\rho(\tau)$ be the state of the chain at time τ. Furthermore, assume that $\lambda_{JI} = \mathrm{O}(1)$ and λ_{IH} is at most $\mathrm{O}(\epsilon)$ and $\Delta \mathrm{t} \succ \mathrm{O}(1)$. Then, if $\rho(t) \neq \mathrm{I}$, we have

$$\lim_{\epsilon \to 0} \frac{E\left[\tilde{\varsigma}(\Delta t)\right] - E\left[\varsigma(\Delta t)\right]}{E\left[\varsigma(\Delta t)\right]} = 0 \tag{4.254}$$

and if $O(\lambda_{IH}^{-1}) \prec \Delta t$, then (4.254) holds without the restriction on $\rho(t)$.

<u>Proof</u>

We can expand the expression on the left side of (4.254) as

$$\begin{aligned} \frac{E\left[\tilde{\varsigma}(\Delta t)\right] - E\left[\varsigma(\Delta t)\right]}{E\left[\varsigma(\Delta t)\right]} &= \frac{E\left[\eta_{IH}(t+\Delta t) - \eta_{IH}(t)\right] - E\left[\eta_{JI}(t+\Delta t) - \eta_{JI}(t)\right]}{E\left[\eta_{JI}(t+\Delta t) - \eta_{JI}(t)\right]} \\ &= \frac{E\left[\eta_{IH}(t+\Delta t) - \eta_{JI}(t+\Delta t)\right]}{E\left[\eta_{JI}(t+\Delta t) - \eta_{JI}(t)\right]}. \end{aligned} \tag{4.255}$$

We consider two cases: $\Delta t = o(\lambda_{IH}^{-1})$ and $\Delta t \succ O(\lambda_{IH}^{-1})$.

<u>Case 1: $\Delta t = o(\lambda_{IH}^{-1})$</u>

We can apply Lemma 1 here to show that only the first transition need be included in the expected values. [1] Therefore since the state at time t might not be

[1] To see why this is true, we can scale time using $\tau = t\ \lambda_{IH}$, yielding $\Delta\tau = \mathrm{o}(1)$ and a new transition rate $\bar{\lambda}_{IH} = 1$.

H, we can obtain an upper bound on the numerator of (4.255) if the initial state is not state I ($\rho(t) \neq I$),

$$\begin{aligned} E\left[\eta_{IH}(t+\Delta t) - \eta_{JI}(t+\Delta t)\right] &< Pr\{\rho(t+\Delta t) = I \mid \rho(t) = H\} \\ &= \int_0^{\Delta t} \lambda_{IH} e^{-\lambda_{IH}\tau} d\tau \left(\int_{\Delta t - \tau}^{\infty} \lambda_{JI} e^{-\lambda_{JI} t} dt \right) \\ &= \frac{\lambda_{IH}}{\lambda_{JI} - \lambda_{IH}} e^{-\lambda_{JI}\Delta t} \left(e^{(\lambda_{JI} - \lambda_{IH})\Delta t} - 1 \right) \\ &\leq \frac{\lambda_{IH}}{\lambda_{JI} - \lambda_{IH}} \left(e^{-\lambda_{IH}\Delta t} \right) \\ &\leq \frac{\lambda_{IH}}{\lambda_{JI} - \lambda_{IH}} \\ &= O\left(\lambda_{IH}^{-1} \right). \qquad (4.256) \end{aligned}$$

Now we must try to obtain a lower bound on the magnitude of the denominator of (4.255). We proceed by noting that the expected value of the denominator is greater than if there were no chance of a transition into state H after time t. i.e. The smallest that it can be is the value obtained using the probability that $\rho(t) =$ H at the start of the interval to calculate an expected value. Hence

$$E\left[\eta_{JI}(t+\Delta t) - \eta_{JI}(t)\right] \geq Pr\{\eta(t+\Delta t) \neq H, I \mid \rho(t) = H\} Pr\{\rho(t) = H\}, \qquad (4.257)$$

but if $\rho(t)$ is H, then

$$\begin{aligned} Pr\{\rho(t+\Delta t) \neq H, I\} &= \int_0^{\Delta t} \lambda_{IH} e^{-\lambda_{IH}\tau} d\tau \left(\int_0^{\Delta t - \tau} \lambda_{JI} e^{-\lambda_{JI} t} dt \right) \\ &= 1 - e^{-\lambda_{IH}\Delta t} - \frac{\lambda_{IH} e^{-\lambda_{JI}\Delta t}}{\lambda_{JI} - \lambda_{IH}} \left(e^{(\lambda_{JI} - \lambda_{IH})\Delta t} - 1 \right) \\ &\geq 1 - \left(1 - \lambda_{IH}\Delta t + \frac{\lambda_{IH}^2 \Delta t^2}{2} \right) \\ &\quad - \frac{\lambda_{IH}}{\lambda_{JI} - \lambda_{IH}} \left(1 - \lambda_{IH}\Delta t + \frac{(\lambda_{IH}\Delta t)^2}{2} \right) \\ &\geq \lambda_{IH}\Delta t - \frac{\lambda_{IH}^2 \Delta t}{\lambda_{JI} - \lambda_{IH}} - \frac{\lambda_{IH}}{\lambda_{JI} - \lambda_{IH}} - O(\lambda_{IH}^3 \Delta t^2) \\ &= O(\lambda_{IH}\Delta t) \qquad (4.258) \end{aligned}$$

Using the fact that $Pr\{\rho(t) = H\} = O(1)$, we now know that

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{E\left[\tilde{\varsigma}(\Delta t)\right] - E\left[\varsigma(\Delta t)\right]}{E\left[\varsigma(\Delta t)\right]} &\preceq \lim_{\epsilon\to 0} O\left(\frac{\lambda_{IH}}{\lambda_{IH}\Delta t}\right)\\ &= \lim_{\epsilon\to 0} O\left(\frac{1}{\Delta t}\right)\\ &= 0 \qquad (4.259)\end{aligned}$$

since $\Delta t \succ O(1)$.

Case 2: $\Delta t \succ O(\lambda_{IH}^{-1})$

In this case we scale time using

$$\tau = t\lambda_{IH} \qquad (4.260)$$

to obtain $\bar{\lambda}_{IH} = 1$ and $\Delta\tau \succ O(1)$. Therefore we can apply Lemma 3 to obtain

$$\lim_{\epsilon\to 0} E\left[\tilde{\varsigma}(\Delta\tau)\right] = \bar{\lambda}_{IH}\pi_{H|m}^{(k)}. \qquad (4.261)$$

However, state H has an O(1) ergodic probability and therefore

$$\begin{aligned}E\left[\tilde{\varsigma}(\Delta\tau)\right] &= O(\bar{\lambda}_{IH})\\ &= O(1) \qquad (4.262)\end{aligned}$$

which yields

$$\begin{aligned}E\left[\eta_{IH}(\tau+\Delta\tau) - \eta_{IH}(\tau)\right] &= E\left[\tilde{\varsigma}(\Delta\tau)\Delta\tau\right]\\ &= \bar{\lambda}_{IH}\Delta\tau\\ &\succ O(1). \qquad (4.263)\end{aligned}$$

If we use the fact that

$$E\left[\eta_{JI}(\tau+\Delta\tau) - \eta_{JI}(\tau)\right] - E\left[\eta_{IH}(\tau+\Delta\tau) - \eta_{IH}(\tau)\right] \leq 1, \qquad (4.264)$$

always, we can write

$$\begin{aligned}\frac{E\left[\tilde{\varsigma}(\Delta\tau)\right]-E\left[\varsigma(\Delta\tau)\right]}{E\left[\varsigma(\Delta\tau)\right]} &= \frac{E\left[\eta_{JI}(\tau+\Delta\tau)-\eta_{JI}(\tau)\right]-E\left[\eta_{IH}(\tau+\Delta\tau)-\eta_{IH}(\tau)\right]}{E\left[\eta_{JI}(\tau+\Delta\tau)-\eta_{JI}(\tau)\right]} \\ &= o(1) \qquad (4.265)\end{aligned}$$

and therefore

$$\lim_{\epsilon\to 0}\frac{E\left[\tilde{\varsigma}(\Delta t)\right]-E\left[\varsigma(\Delta t)\right]}{E\left[\varsigma(\Delta t)\right]}=0. \qquad (4.266)$$

It follows from this result that equation (4.252) is true. Now that we have this result, we may apply it to each of the sequences of transitions so that

$$\lim_{\epsilon\to 0}\frac{E\left[\frac{\eta_{1A,3}(t+\Delta t)-\eta_{1A,3}(t)}{\Delta t}\right]-E\left[\frac{\eta_{3,1A}(t+\Delta t)-\eta_{3,1A}(t)}{\Delta t}\right]}{E\left[\frac{\eta_{3,1A}(t+\Delta t)-\eta_{3,1A}(t)}{\Delta t}\right]}=0 \qquad (4.267)$$

and

$$\lim_{\epsilon\to 0}\frac{E\left[\frac{\eta_{1B,2}(t+\Delta t)-\eta_{1B,2}(t)}{\Delta t}\right]-E\left[\frac{\eta_{3,1B}(t+\Delta t)-\eta_{3,1B}(t)}{\Delta t}\right]}{E\left[\frac{\eta_{3,1B}(t+\Delta t)-\eta_{3,1B}(t)}{\Delta t}\right]}=0. \qquad (4.268)$$

Hence the frequency obtained for appropriate transitions from states 2 and 3 (which have O(1) ergodic probability) is asymptotically equal to the required expected frequency for the state 1 to state 3 transitions. Therefore, for this example, we see that the state splitting and subsequent transition substitution provides a valid interpretation for the calculations of our algorithm.

Chapter 5

Decomposition of the FMS Model

5.1 Introduction

In Chapter 3, a model for a small flexible manufacturing system was developed. The model is intended to be simpler than a realistic system to minimize calculations, while still demonstrating features that can be expected in a real system. This chapter describes the time scale effects that can be expected and then uses the techniques of Chapter 4 to quantify these effects for the model. The result is that the original high-order model is replaced by several lower dimensional versions at different time scales. Once the decomposition calculations are complete, the average frequencies of part production are calculated and these results are in turn used to calculate capacities for the production of the system.

5.2 Preparation for the Decomposition

In Chapter 3, a Markov chain model of the system was developed. The corresponding transition rate matrix for the chain was then formed (equations 3.10 to 3.25).

The elements of the matrix correspond to the rates of the events that occur in the system: failures, repairs, setups, operations and decisions. These rates are assumed to have four different orders of magnitudes. The failure and repair rates are assumed to have the smallest magnitude, $O(\epsilon^3)$, where ϵ is a small positive number. The setups are assumed to occur with a frequency that is $O(\epsilon^2)$, operation frequencies are $O(\epsilon^1)$ and decisions, the fastest events, take place at a rate that is O(1). Since the rate differences are already explicit in the transition matrix through the parameter ϵ, we can directly apply the techniques of Chapter 4 to the model.

Note here, that since we have the transition rate matrix $\underline{A}(\epsilon)$, we have all that is required to perform the time scale decomposition algorithm using Rohlicek's algorithmic approach described in Section 4.3. If knowledge regarding event frequencies is required, the algorithm of Section 4.5.3 can subsequently be implemented.

Before actually carrying out the decomposition, we provide a qualitative description of what the analysis will reveal. This description is based partially on knowledge of the results that we will obtain, but is intended only as an aid in understanding the results for this example and is in no way required to implement the decomposition. State transition diagrams are provided in the preliminary discussion as reference aids. These will correspond exactly with the qualitative results except that we will label the transitions rates following the decomposition.

We start by considering observations of the chain over a very long time interval. For this time interval, all of the transitions in the chain will occur many times. The result is that all of the events in the system are blurred and individual transitions cannot be distinguished. Therefore an appropriate chain for modeling the system at this time scale consists of exactly one aggregate state. This corresponds to the system being in a complete steady state. Hence we cannot distinguish any individual

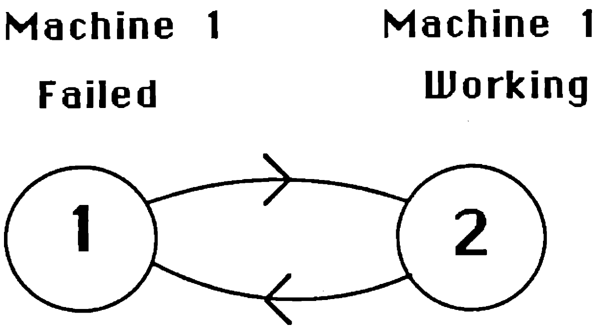


Figure 5.1: Expected Model at Time Scale of Failure Events

changes in the FMS, so only one state is required in the model.

If we look at the system over a somewhat shorter time horizon, we obtain a situation in which the individual transitions between failure and repair modes of the system are observed. Since the failure and repair events are the slowest events that occur in the system, all other events will occur much more frequently and hence will appear as a continuous stream of transitions. We expect a model at this time scale to consist of exactly two states, one for machine 1 working and one for machine 1 failed. This model is depicted in Figure 5.1.

While the system is in each of the aggregate failure states, the faster events such as setup initiations and completions and operation related events continue to occur, but take place so frequently that individual events become blurred together.

If we decrease the time interval of observation even further, the probability of observing failure-related events becomes very small, but we will distinguish individual setup events for machine 2. The dynamics for the system that we expect to see at this time scale are those which are generated by the set-up events of the system, and are shown in a state transition diagram in Figure 5.2. The transition behavior shown in Figure 5.2 will exist for both of the failure modes of the system.

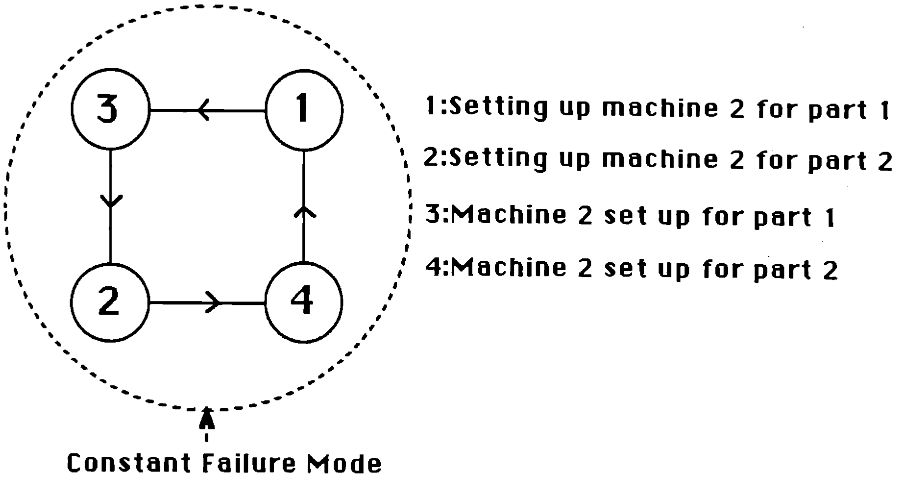

3
1
2
4
1:Setting up machine 2 for part 1
2:Setting up machine 2 for part 2
3:Machine 2 set up for part 1
4:Machine 2 set up for part 2
Constant Failure Mode

Figure 5.2: System dynamics at time scale of Setup events.

We can now repeat the process by examining the system at a shorter time scale. Over time scales which are shorter than those examined above, the probability of observing failure or set-up events will become very small, but individual events associated with operations will be observable. In this case, unlike the set-up case, the dynamics of the system will differ depending on the current failure or set-up mode of the system. This is because the system cannot enter some operational states if machine 1 has failed or machine 2 is in the process of being set up. In Figure 5.3, the most complex operational dynamics possible for our system are depicted, when machine 1 is working and machine 2 is set up for part 1. The coordinate pairs (a,b) that label each of the states represent the operational modes of machine 1 and machine 2. The possible values for a or b are: 0 for the idle mode and 1 or 2 for operational modes 1 or 2 respectively.

By decreasing the time interval of observation again, we will observe the fastest events in the system, the decision events. All events described so far will have a

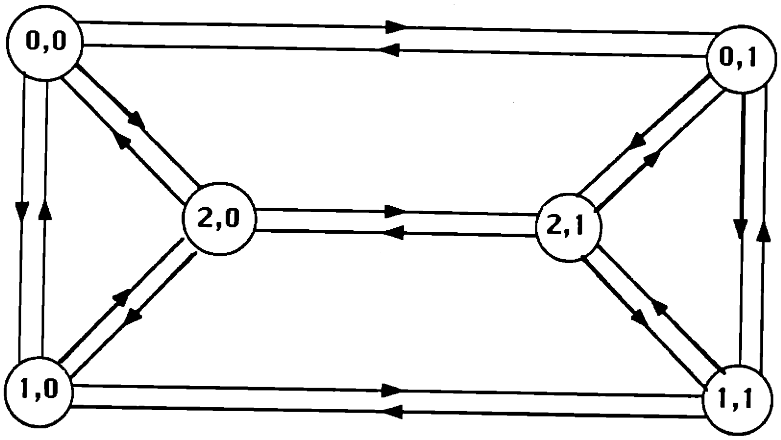


Figure 5.3: Operational Dynamics

very small probability of occuring, because they occur at much slower rates than the decisions. The decision dynamics will vary for different failure and set-up modes because certain machines may not be capable of operating on certain parts. The decision dynamics are depicted in Figure 5.4 for the most complex case of no current failures or setups. The pairs of symbols used to label the states are the same for this figure as in Figure 5.3, with 0,1 and 2 representing the operational modes for machines 1 and 2. In this figure there is the additional symbol D which indicates that the machine is idle and a decision is being made regarding the next operational mode to enter. The figure shows only transitions from the decision states to the operational states and no transitions in the reverse direction. This is because those transitions would correspond to operation completions, which occur with very low probability at this time scale.

In summary we see that the time scale decomposition should generate models on five different time scales (counting the extremely long time scale where all events

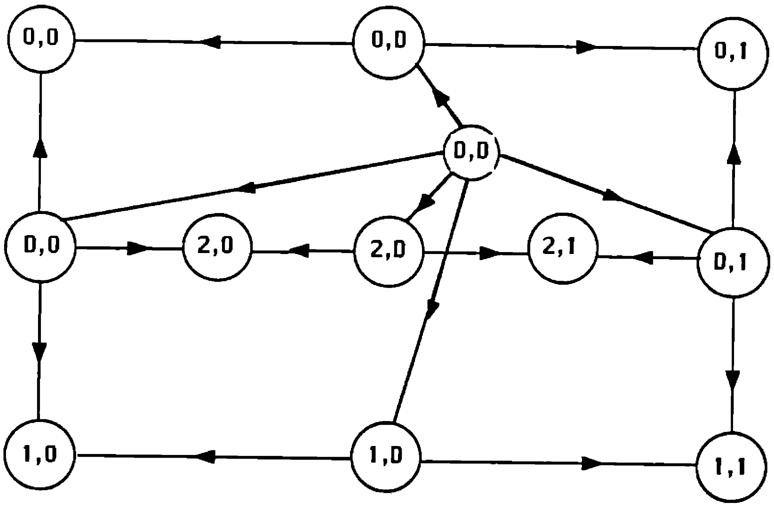


Figure 5.4: Decision Dynamics

are blurred.) We expect that the models at each of these time scales will be of significantly lower dimension than the original model, simply because each model only contains transitions relating to a single event or type of event in the system.

5.3 The Time Scale Decomposition

Given that we have a qualitative idea of what to expect from the time scale decomposition, we can proceed to verify these expectations using the techniques of Chapter 4. The model of Chapter 3 provides a transition rate matrix with elements corresponding to rates at which events occur. These rates, which are denoted by $\lambda_{ij}(\alpha,\sigma)$, $\epsilon\tau_{ij}^{-1}$, $\epsilon^2 f_s(\alpha,\sigma)$, $\epsilon^2 s^{-1}$, $\epsilon^3 p$ and $\epsilon^3 r$, are fundamental quantities that are determined by the physical features such as failure characteristics of the system that we are modeling. When we perform the decomposition, a number of new rates will be defined for the aggregate models which are combinations of the fundamental quantities. To maintain consistency with other parts of the analysis, these quantities will be defined with a superscript indicating the level at which these derived rates are defined. For example, $p^{(3)}$ is the failure rate determined at level 3. Strictly speaking, the fundamental quantities whould be written with a superscript (0) to indicate that they are defined at level 0. However, the superscript is dropped for the fundamental quantities so that they may be easily distinguished from the derived rates.

Noting that the decision rates are O(1) in the original transition rate matrix defined in Chapter 3, the initial model describes the system at a time scale commensurate with the mean time for a decision to be made. We start at this time scale and work toward the slower models since this is the order dictated by the techniques of Chapter 4. This is the opposite order to that taken for the qualitative description of Section 5.2, but the results must be the same. Also, we reference time scales with indices 0,1,2..., where 0 is the shortest time scale while in Gershwin [10], 1 is the longest time scale and the time scales become shorter as the index increases.

We start by noting that the transition matrix has dimension 40 by 40 and therefore the state partitioning techniques of Sections 3.5 and 4.A should be used. Our choice of partitioning criteria is the failure and set-up modes of the system. Since there are two possible failure modes and four possible set-up modes, this yields a total of eight subspaces for the system. Noting that transition rates between the subspaces are all $O(\epsilon^2)$ and $O(\epsilon^3)$, the simplified methods for calculations that were presented in Section 4.A may be applied. A more detailed justification for this partitioning is given in Appendix 5.A. The details of the calculations and results are also provided Appendix 5.A; only those results necessary to demonstrate agreement with the previous section are presented here.

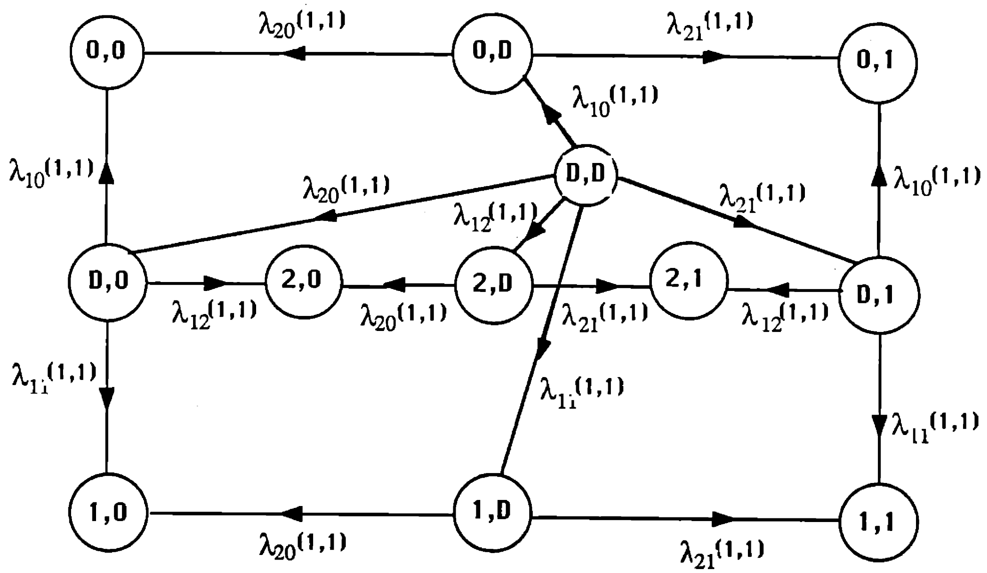


Figure 5.5: State transition diagram at fastest time scale ($\epsilon = 0$)
(Machine 1 in working mode, Machine 2 set up for part 1)

We start at the fastest time scale and determine the behavior of the system as $\epsilon \rightarrow 0$, so that all slower events will occur with a probability that approaches 0. This behavior is predicted by the transition matrix since the only rates that do not vanish are those at which decisions are made. If we draw the state transition diagram for the system at this fastest time scale, we obtain Figure 5.5. (The state labels are defined as in Section 5.2, with operational modes of 0,1 and 2 and a decision mode of D for each machine).

The transition rate matrix for the chain at the next time scale is obtained by performing an aggregation as described in Section 4.3. The details of this calculation are provided in the appendix to this chapter. The matrix that is obtained provides a description of the dynamics of the chain over time intervals that are $O(\epsilon^{-1})$. Therefore, ϵt is $O(1)$ and the transitions which have $O(\epsilon)$ rates will occur with an $O(1)$ probability. The transitions corresponding to decision completions, have $O(1)$ rates and therefore will occur so quickly that they will not be distinguishable at

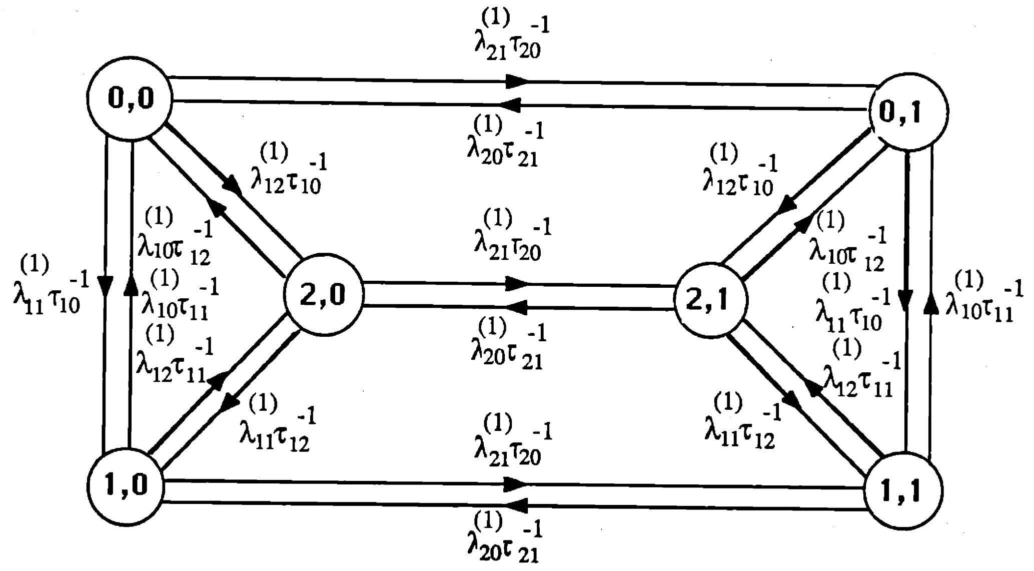

$\lambda^{(1)}_{21}\tau^{-1}_{20}$
0,0
0,1
$\lambda^{(1)}_{20}\tau^{-1}_{21}$
$\lambda^{(1)}_{12}\tau^{-1}_{10}$
$\lambda^{(1)}_{12}\tau^{-1}_{10}$
$\lambda^{(1)}_{10}\tau^{-1}_{12}$
$\lambda^{(1)}_{21}\tau^{-1}_{20}$
$\lambda^{(1)}_{10}\tau^{-1}_{12}$
$\lambda^{(1)}_{11}\tau^{-1}_{10}$
$\lambda^{(1)}_{10}\tau^{-1}_{11}$
2,0
2,1
$\lambda^{(1)}_{11}\tau^{-1}_{10}$
$\lambda^{(1)}_{10}\tau^{-1}_{11}$
$\lambda^{(1)}_{12}\tau^{-1}_{11}$
$\lambda^{(1)}_{20}\tau^{-1}_{21}$
$\lambda^{(1)}_{12}\tau^{-1}_{11}$
$\lambda^{(1)}_{11}\tau^{-1}_{12}$
$\lambda^{(1)}_{21}\tau^{-1}_{20}$
$\lambda^{(1)}_{11}\tau^{-1}_{12}$
1,0
1,1
$\lambda^{(1)}_{20}\tau^{-1}_{21}$

Figure 5.6: State transition diagram at second time scale.

this time scale. Figure 5.6 shows the state transition diagram which represents the dynamics for one of the ergodic classes at this time scale. The terms $\lambda^{(1)}_{ij}$ that are used in the expressions for the transition rates in the diagram are defined by

$$\lambda^{(1)}_{ij} = \frac{\lambda_{ij}(1,1)}{\sum_k \lambda_{ik}(1,1)}$$

This particular ergodic class contains all of the states for which machine 1 is working and machine 2 is set up to operate on part 1. This corresponds to the diagram in Figure 5.4; however, we should note that this diagram as well as those in Figures 5.6 to 5.8 are generated completely from the transition rate matrix as obtained from the time scale decomposition.

We may now repeat the aggregation procedure using the algorithm of Section 4.3.2. The result is a model that is valid over time intervals of length $O(\epsilon^{-2})$. The state transition diagram for the entire process at this time scale is shown in Figure

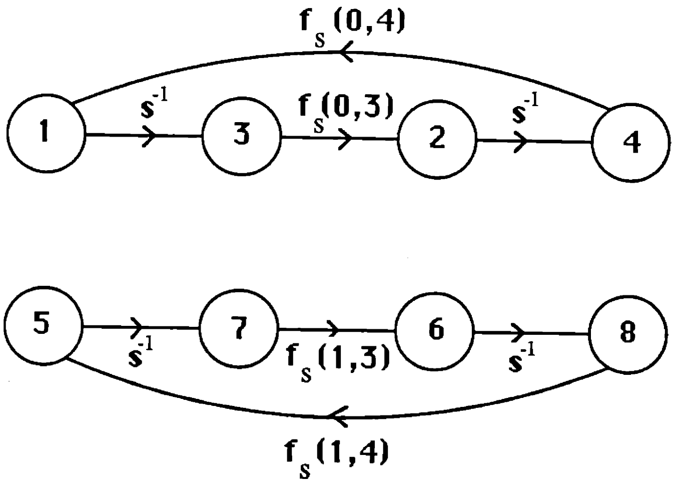


Figure 5.7: State transition diagram at third time scale

5.7. The expressions $f_s(i,j)$ in that figure correspond to the rate at which setups are initiated, given that the failure component is in mode i and the set-up component is in mode j. The expressions $p^{(2)}(1,j)$ are given by

$$p^{(2)}(1,j) = p\left(\frac{\lambda_{11}(1,j)\tau_{11} + \lambda_{12}(1,j)\tau_{12}}{\sum_k \lambda_{1k}(1,j)\ \tau_{1k}}\right).$$

This expression generates a reduced failure rate that compensates for the fact that failures only occur when a machine is operating, but the machine may not be operating 100% of the time. The corresponding transition rate matrix is derived in the appendix. Each of the 8 states corresponds to a unique combination of modes for the failure and set-up components of the state. Table 5.1 provides a list of the state labels and the associated modes for the failure and set-up components. At this time scale, we observe the system over time intervals which are $O(\epsilon^{-2})$ and hence $\epsilon^2 t$=O(1). Therefore, the transitions with $O(\epsilon^2)$ rates, corresponding to setup related events, now have an O(1) probability of occurring, while the operation related events that we saw at the previous time scale are blurred away. The transition

State	Failure Mode (Machine 1)	Setup mode (Machine 2)
1	Failed	Setting up for part 1
2	Failed	Setting up for part 2
3	Failed	Set up for part 1
4	Failed	Set up for part 2
5	Operating	Setting up for part 1
6	Operating	Setting up for part 2
7	Operating	Set up for part 1
8	Operating	Set up for part 2

Table 5.1: Description of The Eight Aggregate States at the Third Time Scale

diagram shown in Figure 5.7 supports this, since the transitions are due to setup initiations (with rate $f_s(i,j)$) and setup completions (with rate s^{-1}).

By repeating the aggregation calculations of Section 4.3.2 once again, we obtain a transition matrix of dimension 2, which is given by Equation (5.63) in the appendix. The state transition diagram corresponding to this transition matrix is shown in Figure 5.8. The rate $p^{(3)}$ in the diagram is given by

$$p^{(3)} = \sum_{j=1}^{4} a_{j2} p^{(2)}(1,j),$$

where $p^{(2)}(i,j)$ was defined in Figure 5.7. We also introduce the parameter a_{ij} which is the fraction of time spent in set-up mode i, while the system is in failure mode j. For example b_{10} is the fraction of time spent setting up machine 2 for part 1 (mode 1), while machine 1 is failed (mode 0), and is given by

$$a_{10} = \frac{s f_s(0,3)\ f_s(0,4)}{2s f_s(0,3)\ f_s(0,4)\ +\ f_s(0,3)\ +\ f_s(0,4)}.$$

Expressions for a_{ji} for each i and j are provided in the appendix. The resulting expression for $p^{(3)}$ is therefore a weighted sum of the failure rates for each set-up mode of the previous time scale, with weights that are equal to the fraction of time spent in each of those modes.

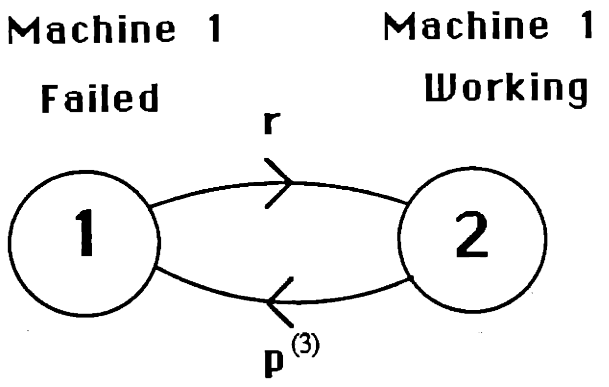


Figure 5.8: State transition diagram at fourth time scale

At this time scale, we observe transitions which are $O(\epsilon^{-3})$ for reasons similar to those provided for shorter time scales. The transitions which are $O(\epsilon^{-3})$ are related to the failure and repair events. Examining Figure 5.8, we see that there are two states as we predicted in Section 5.3 and the transition rates are functions of the repair and failure rates.

Finally, we can repeat the aggregation procedure one last time. After this last aggregation we obtain a chain with a single aggregate state, as demonstrated in the appendix. The reason is that all of the elements in the transition matrix at the fourth time scale were $O(1)$ so that the system does not display behavior at additional time scales. Therefore, by increasing the interval of observation this last time, we blur out all remaining transitions in the chain.

5.4 Frequencies of Operation Completion

Having completed the time scale decomposition of our simple model, it is possible to calculate the expected frequency of various events represented in the model. In the case of manufacturing systems, we are typically interested in the rate at which parts are produced so that we can match demand. For our simple Markov Chain model, we note that completion of an operation is equivalent to the production of a part because each part undergoes only a single operation. We may therefore calculate expected frequencies of production for each part and for each machine by determining the expected frequencies of transitions out of states in which operations are being performed. These results are not directly required if we are only interested in the total production frequency of a part (only their sum is required). They will, however, be needed if for example we wish to calculate the fraction of time a machine is in use.

It is important to note that production frequencies can be calculated, and in general will be different, for each of the individual states at each time scale. For example, there are only two aggregate states at time scale 3 (based on time intervals commensurate with the mean time between failures). These states correspond to machine 1 being in either failed or working condition. Therefore, at that time scale, there are two production rates for either part, one corresponding to each of the two possible aggregate states. In addition, we obtain the production frequencies for individual machines in our preliminary calculations. We procede now in the manner of Section 5.2, providing a qualitative description of the results that we expect.

Starting with time scale 0 (the fastest time scale) we are observing the system over time intervals with lengths that are $o(1)$. However, since operation completions correspond to transitions with rates $O(\epsilon)$, the probability that a transition corresponding to an operation completion will occur is $o(\epsilon)$, which becomes very small for small ϵ. Therefore, the probability that two operations will be completed is very small and may be ignored. The result is that the expected frequency of operation completions at this time scale is equal to the Markovian transition rate for the associated transitions if an operation is in progress. If an operation is not in progress, the expected frequency will be zero (the probability of both initiating and completing an operation will be very small).

At the next time scale, the intervals of observation are $o(\epsilon^{-1})$. The probability of an operation completion during such an interval is $o(1)$, which still becomes small as $\epsilon \rightarrow 0$. We may therefore repeat the arguments from time scale 0 to show that the expected frequency of part completions should once again be the Markovian transition rates, because the probability of completing more than one operation is sufficiently small that it may be ignored.

When the model at the next time scale is considered, the number of states is decreased to eight. At this time scale, multiple operations may be completed by a single machine. Since some of these operations may be different, the different operation types will compete for machine time. Therefore the production frequency of a part on a machine will be affected by its own decision parameters and operation completion times, as well as those of other operations on the machine. Since these parameters may depend on the failure and set-up modes of the system, the production frequencies will also be a function of these modes.

For example, if we consider the production of part 1 on machine 1 at this time scale, the machine will perform many part 1 and part 2 operations and spend some time idle. Therefore the frequency of part 1 production will depend on the parameters that determine which operation related activity is initiated next – i.e. operation on part 1, operation on part 2 or no operation. We also note that the decision parameters are a function of the failure and set-up modes and therefore the production frequencies will also be functions of these modes.

When we extend the time horizon again, we obtain the two state aggregate model based on the failure modes of the system. Therefore we obtain production frequencies which are different for the two failure modes of the system. The frequency in each of the states will depend on the set-up initiation and completion rates, $f_s(i,j)$ and s^{-1} because these rates determine the fraction of time that machine 2 is capable of operating on each of the parts.

Finally, at a time scale which is much longer than the mean time between the slowest transitions, we obtain the single state model described in Section 5.3. The production frequency at this time scale will depend on the amount of time that machine 1 is in working order. Therefore we expect that the production frequencies at level 4 will be a function of the production frequencies at level 3, as well as the failure and repair rates.

5.5 Calculation of Production Frequencies

In order to facilitate the description of the expected production frequencies, we introduce the following notation.

Definition:

The symbol ${u_{ij}}^{(k)}(\alpha,\sigma)$ will be used to denote the expected frequency of part j production on machine i. The superscript k indicates the time scale index, where level 0 is associated with the shortest time scale and level 4 is associated with the longest time scale of the system. The arguments α and σ correspond to the modes of the failure and setup components. Let $m^{(k)}$ be an aggregate state in the model of the system at time scale k such that the failure mode of the system is α and the set-up mode is σ, while $a \prec b$ indicates that $\lim_{\epsilon\to 0}\frac{a}{b}=0$. Furthermore let $\eta(t)$ be the number of type j parts produced by machine i up to time t. Then we define

$$u_{ij}^{(k)}(\alpha,\sigma) \equiv E\left[\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t} \mid \rho(t) = m^{(k)}, O(\epsilon^{-k+1}) \prec \Delta t = o(\epsilon^{-k})\right]. \quad (5.1)$$

For example, the expression $u_{45}^{(3)}(1,2)$ refers to the frequency of production of part 4 on machine 5. The time interval of observation satisfies $O(\epsilon^{-2}) \prec \Delta t = o(\epsilon^{-3})$, while the failure component is in mode 1 and the set-up component is in mode 2. This notation is consistent with that of Gershwin [10], with the exception that the ordering of the index for the time scale (k) is reversed from that in [10]. This dependence on the failure and set-up modes is not meaningful at all time scales. For example, at level 3 we have 2 states, defined by the failure mode only, because the set-up modes have been aggregated away. Therefore, for level 3 rates, we drop the σ argument. For similar reasons, both of the arguments are dropped at level 4.

We proceed to describe the quantitative results obtained using the techniques of Section 4.5. The numerical aspects are provided in the appendix, with only those results that are important for a conceptual understanding repeated here.

Starting at the fastest time scale, we see that the matrix of expected production frequencies are either zero or $O(\epsilon)$. The non-zero frequencies are of the form $\epsilon\tau_{ij}^{-1}$ which is simply the inverse of the mean time to completion of an operation. This expression is the Markovian transition rate as predicted.

At the next time scale, level 1, the expected frequencies are zero for the states in which the machine of interest is not operating on the part of interest. For the remaining states, the expected production frequency is just the inverse of the mean time to completion or $\epsilon\tau_{ij}^{-1}$. The only difference from the previous time scale is that we have fewer states, because the states for which decisions are being made have been aggregated away.

At the time scale of level 2, we calculate the frequencies over intervals which are $O(\epsilon^{-2})$. The production frequencies at this time scale are provided in the appendix and are of the form

$$u_{ij}^{(2)}(\alpha,\sigma) = \frac{\lambda_{ij}(\alpha,\sigma)}{\sum_k \lambda_{ik}(\alpha,\sigma)\tau_{ik}}, \tag{5.2}$$

for the states in which machine i is capable of production and zero otherwise. For example, consider the aggregate state at this time scale which corresponds to machine 2 being capable of working on part 2, but not part 1. The elements in the transition frequency matrix given in (5.70) of the appendix, which correspond to the terms $u_{12}^{(2)}(\alpha,4)$, are zero (there is a zero rate of production of part 1 on machine 2, while it is setup to work on part 2). The frequencies were calculated using the algorithm of Section 4.5.3. The expression in equation (5.2) corresponds physically to the rate at which operations are completed given an operation is being performed, multiplied by the fraction of time that the machine operates on a type j part.

Note that the denominator of the expression in (5.2) contains parameters corresponding to all of the operations on machine i. This comes from the expression for the fraction of time spent operating on type j parts and is a manifestation of the competitive aspect of the operations for machine time. To see this we need only recognize that changes in the time required by machine 1 to complete an operation on a type 2 part (τ_{12}) will effect the production frequency for part 1 ($u_{11}^{(2)}(\alpha,\sigma)$). Finally, the product of the expected frequencies and the length of the time interval satisifies $\left[\underline{Q}^{(2)}\right]_j \Delta t \succ O(1)$ (for those elements $\left[\underline{Q}^{(k)}\right]_{ji} \neq 0$). and therefore the observed frequencies will equal the calculated frequencies with probability 1 as $\epsilon \to 0$.

At the next time scale, corresponding to level 3, we obtain the operation completion frequencies in equation (5.71) of the appendix. The expressions for the frequencies have one of three forms depending on which machine we are considering and whether or not machine 1 is in working order. First we note that $u_{11}^{(3)}(0)$

and $u_{12}^{(3)}(0)$ are equal to zero. This makes intuitive sense because machine 1 cannot produce either part 1 or part 2 when it is failed. When machine 1 is in working order, we obtain production frequencies of the form

$$u_{1i}^{(3)}(1) = \sum_{j=1}^{4} \left(\frac{\lambda_{1i}(1,j)}{\sum_{k=0}^{2} \lambda_{1k}(1,j)\ \tau_{1k}} a_{j1} \right) \tag{5.3}$$

which is just a linear combination of the level 2 frequencies, with weights given by the terms a_{j1}. These quantities represent the fraction of the time spent in each setup mode and are functions of the setup initiation and completion rates, $f_s(\alpha,\sigma)$ and s^{-1}. For example, a_{11} is the fraction of the time spent in set-up mode 1, given that the machine is working (failure component in mode 1). It is given by

$$a_{11} = \frac{s f_s(0,3)\ f_s(0,4)}{2s f_s(0,3)\ f_s(0,4)\ +\ f_s(0,3)\ +\ f_s(0,4)}. \tag{5.4}$$

Expressions for the remainder of the a_{ij} terms are given in the appendix by equations (5.59) to (5.61). Finally, we have the production frequencies for machine 2 at level 3, which have a similar form to the expression in (5.3) except that in this case the expression contains only the frequency corresponding to one of the set-up modes. This is easily explained if we consider the frequency of part 1 production on machine 2. This frequency will only be non-zero when machine 2 is set up to operate on part 1 (mode 3). Therefore we obtain

$$u_{21}^{(3)}(i,3) = \left(\frac{\lambda_{21}(i,3)}{\sum_{k=0}^{1} \lambda_{2k}(i,3)\ \tau_{2k}} \right) a_{3i}. \tag{5.5}$$

The expressions for the production frequencies of part 2 are similar, and are given explicitly by equation (5.71) in the appendix. Once again, for $\left[\underline{Q}^{(k)}\right]_{ij} \neq 0$, $\left[\underline{Q}^{(3)}\right]_{ij} \Delta t \succ O(1)$ and therefore Theorem 3 of Section 4.5.6 applies for the sample path and expected frequencies.

At the longest time scale, level 4, we have the rates given by (5.72) in the appendix. In this case, the frequencies are linear combinations of the level 3 rates.

The weightings in this combination are functions of the failure and repair rates at level 3, because these rates determine the fraction of time that the machine is failed and the fraction of time it is in working order. For example, suppose we consider the production frequency for type 1 parts on machine 2. Examining the appropriate element of equation (5.72), we see that

$$u_{21}^{(4)} = \frac{p^{(3)}}{r + p^{(3)}}\, u_{21}^{(3)}(0) \;+\; \frac{r}{r + p^{(3)}}\, u_{21}^{(3)}(1), \qquad (5.6)$$

which is just the fraction of time spent in the failed condition multiplied by the production frequency when failed plus the fraction of time that machine 1 is in working order multiplied by the production frequency for that condition. Note again that Theorem 3 applies for the sample path frequencies.

It is interesting to note the form of the calculations that we are performing here. Specifically, we are applying the equation $\underline{Q}^{(k)} = \underline{Q}^{(k-1)}\underline{U}^{(k-1)}$ from Section 4.5.3. The right hand side of this expression is the level k-1 production frequencies multiplied by the ergodic probabilities of the level k-1 aggregate states. Therefore we are effectively taking the expected value at time scale k-1. This is essentially the same operation that is performed by equation (11) in Gershwin [10].

5.6 Calculation of Capacity Constraints

In Gershwin [9], the concept of capacity is introduced. The basic idea behind capacity is that there is a maximum number of parts that can be produced per unit time regardless of how many unmachined parts are introduced into the FMS. This information is important because loading parts at a rate which is greater than the capacity will not increase total production, but may slow down the system due to congestion.

Gershwin also introduces the concept of capacities that are dependent on the time scale being considered. For example, suppose that we are considering an FMS with one machine and one part. If failures of the machines are rare compared to operations, then we may consider the system over a short time scale such that a machine failure or repair will occur with very low probability, but many operations will be completed. At this time scale we may define two capacities, one for when the machine is failed (0) and one for when the machine is in working order (maximum operation rate on the machine). If we examine the system over a very long time scale, we will obtain a different capacity. Many machine failures and repairs will occur, and therefore this long run capacity will be a function of both the short run capacity when the machine is operating and the fraction of time that the machine is in working order.

A number of equations are established in [9] for capacities at different levels in the hierarchy (corresponding to different time scales). We will now show that we can obtain these same capacity constraints for our simple model of Chapter 3, using the transition frequency results of Chapter 5. The capacity constraints are generated in a structured manner, so that equivalent results can be systematically generated for more complex systems.

We start by considering $\underline{Q}^{(0)}$ and $\underline{Q}^{(1)}$, the matrices of expected part completion frequencies. These matrices contain only the transition rates for operation completions. For $\underline{Q}^{(0)}$ and $\underline{Q}^{(1)}$, we are considering time intervals which are $o(1)$ and $o(\epsilon^{-1})$ respectively. Therefore the probability of more than one operation completion is very small. Therefore, since we are considering (with high probability) only a single operation completion, the capacity constraint becomes degenerate, and equals the inverse of the mean time to complete that part.

Next consider the matrix $\underline{Q}^{(2)}$. The non zero elements of this matrix are obtained from (5.70) in the appendix as

$$u_{ij}^{(2)}(\alpha,\sigma) = \frac{\lambda_{ij}(\alpha,\sigma)}{\sum_k \lambda_{ik}(\alpha,\sigma)\ \tau_{ik}}. \qquad (5.7)$$

Our next step in generating the capacity constraints is to multiply each frequency $u_{ij}^{(2)}(\alpha,\sigma)$ by τ_{ij} and sum over the operational modes for machine i corresponding to operations on a part. We thereby obtain:

$$\begin{aligned}\sum_{k\neq 0} u_{ij}^{(2)}(\alpha,\sigma)\ \tau_{ij} &= \frac{\sum_{k\neq 0} \lambda_{ik}(\alpha,\sigma)\ \tau_{ik}}{\sum_k \lambda_{ik}(\alpha,\sigma)\ \tau_{ik}} \\ &= 1 - \frac{\lambda_{i0}(\alpha,\sigma)\ \tau_{i0}}{\sum_k \lambda_{ik}(\alpha,\sigma)\ \tau_{ik}} \\ &\leq 1 \qquad (5.8)\end{aligned}$$

because $\lambda_{ik}(\alpha,\sigma),\ \tau_{ik}\ \geq 0$. The expression that is subtracted form the right side of this expression is simply the fraction of time that the machine spends idle. Therefore we obtain

$$\sum_{k=1,2} u_{ij}^{(2)}(\alpha,\sigma)\ \tau_{ij} \leq 1. \qquad (5.9)$$

This is exactly the form of equation (41) in Gershwin [10]. Note that a similar result can be obtained for each machine using the frequencies given in (5.70). We also note that the capacity constraint described by (5.9) is dependent on the failure and setup modes. For example, if we want the capacity for production on machine 1 when machine 1 is failed and machine 2 is set up for part 2, we obtain

$$u_{11}^{(2)}(0,4),\ u_{12}^{(2)}(0,4) = 0, \qquad (5.10)$$

and therefore

$$\sum_{j=1,2} u_{1j}^{(2)}(0,4)\ \tau_{1j} = 0. \qquad (5.11)$$

Proceeding to the next time scale, we obtain another set of frequency equations. From Section 5.5, there are three general forms for the frequencies at this time scale, depending on which machine or part we are considering. For example, consider the capacity restrictions on machine 2. The expressions for the production frequencies for part j are

$$u_{2j}^{(3)}(1) = \left(\frac{\lambda_{2j}(1, j+2)}{\sum_{k=0,1} \lambda_{2k}(1, k+2)\ \tau_{2k}} \right) a_{j+2,1}. \tag{5.12}$$

where $a_{j+2,1}$ is defined as the fraction of time spent in set-up mode j+2 while machine 1 is in working order. The expression $\lambda_{2j}(1, j+2)$ is the decision parameter for machine 2, part j and τ_{2k} is the mean time required to complete operation type k. Repeating our procedure from the previous time scale, we multiply both sides of (5.12) by τ_{2i} and then sum over the possible operations to obtain

$$\begin{aligned}
\sum_{i \neq 0} u_{2i}^{(3)}(1)\tau_{2i} &= \frac{\lambda_{21}(1,3)\ \tau_{21}}{\sum_{k=0,1} \lambda_{2k}(1,3)\tau_{2k}}\ a_{31}\ +\ \frac{\lambda_{22}(1,4)\ \tau_{22}}{\sum_{k=0,2} \lambda_{2k}(1,4)\ \tau_{2k}}\ a_{41} \\
&= \left(1 - \frac{\lambda_{20}(1,3)\ \tau_{20}}{\sum_{k=0,1} \lambda_{2k}(1,3)\ \tau_{2k}}\right) a_{31}\ +\ \left(1 - \frac{\lambda_{20}(1,4)\ \tau_{20}}{\sum_{k=0,2} \lambda_{2k}(1,4)\ \tau_{2k}}\right) a_{41} \\
&\leq\ a_{31}\ +\ a_{41} \\
&=\ 1\ -\ (a_{11}\ +\ a_{21}).
\end{aligned} \tag{5.13}$$

Therefore

$$\sum_{i \neq 0} u_{2i}^{(3)}(1)\tau_{2i}\ +\ a_{11}\ +\ a_{21}\ \leq\ 1 \tag{5.14}$$

where $a_{11}\ +\ a_{21}$ is the fraction of time spent setting up machine 2. Equation (5.14) corresponds to equation (33) of Gershwin [10], with the process of setting up defined as an activity.

Finally we can repeat the procedure at time scale 4 where the production frequencies for machine 1 are

$$u_{1i}^{(4)} = \frac{r}{r + p^{(3)}} \sum_{j=1}^{4} \frac{\lambda_{1i}(1,j)\ a_{j1}}{\sum_{k=0}^{2} \lambda_{1k}(1,j)\ \tau_{1k}}. \tag{5.15}$$

Once again premultiplying by τ_{1i}, summing over the operations and solving in the manner of the previous time scale, we obtain

$$\sum_{i=1,2} u_{1i}^{(4)}\, \tau_{1i} \;+\; \frac{p^{(3)}}{p^{(3)}+r} \;\leq\; 1. \tag{5.16}$$

Intuitively, this corresponds to a reduction in the available capacity for operations on machine 2 due to the time spent by machine 1 in the failed mode. The inequality in (5.16) agrees with equation (32) in [10] with a failure defined as an activity at machine 1.

We see that with a small amount of manipulation, the transition frequencies can be used to yield capacity constraints at each time scale and for each aggregate state at those time scales.

5.7 Summary

In this chapter we have used the techniques of Chapter 4 to obtain reduced order models at various time scales for the simple system introduced in Chapter 3. The techniques enabled us to calculate the frequency of part production for individual machines and the system as a whole. For both the decomposition of the chain into low order models and the calculation of production frequencies, intuitive arguments were presented for the qualitative aspects of results that should be obtained from the calculations and perfect agreement with analytical results was reached.

The three most important aspects of the results are the dimension of the reduced models, the theoretical agreement of the calculated frequencies with those that would actually be observed for a sample path, and the ease with which capacity constraints are calculated. We note that the original model of the system had dimension 40, while the largest model obtained from the time scale decomposition had dimension 12 (fastest time scale, machine 1 operating, machine 2 set up). Secondly, we noted that for time scales 2,3,and 4, the conditions of Section 4.5.6 were met for the production frequencies and the observation intervals. Therefore at the time scales of levels 2 to 4, the results of Chapter 4 indicate that the sample path frequencies will equal the expected value of the frequencies with probability 1 as $\epsilon \to 0$. A special note was made regarding the similarity of the event frequency calculations here and those of Gershwin. Finally, it was shown that the capacity constraints for the FMS can easily be determined using the results from our event frequency calculations.

5.A Appendix

5.A.1 Time Scale Decomposition Calculations

In this appendix we apply the analysis techniques of Chapter 4 to the model of a simple system that was presented in Chapter 3. The calculations and numerical results corresponding to the verbal descriptions of Sections 5.3 and 5.5 are provided.

We start by noting that there are 40 states in the Markov chain model. Therefore the transition rate matrix has dimension 40 x 40 and hence is difficult to manipulate directly. We therefore wish to use a technique for which it is not necessary to work with the entire matrix at one time. We found in the appendix to Chapter 4 that the state space can be partitioned into subspaces so that it is only necessary to deal with smaller submatrices of the transition rate matrix at any given time.

The restriction on the selection of the subspaces is that the transitions between states in different subspaces be at most $O(\epsilon)$. Noting the assumptions of Section 3.6 regarding the magnitude of transition rates, the rates for events associated with the setups and failures are $\mathrm{O}(\epsilon^2)$ and $\mathrm{O}(\epsilon^3)$ respectively. Therefore, if each subspace that we select contains all of the states for a particular combination of failure and set-up modes, the non-zero transition rates between subspaces will be $\mathrm{O}(\epsilon^2)$ or $\mathrm{O}(\epsilon^3)$. Therefore, we will have a subspace corresponding to machine 1 failed and machine 2 setting up for part 1, a subspace for machine 1 working and machine 2 setup for part 2 and so forth, yielding a total of eight subspaces. Table 5.2 lists the modal combinations for each of the eight subspaces.

Subspace	Failure Mode	Set-up Mode
1	Failed	Setting up for Part 1
2	Failed	Setting up for Part 2
3	Failed	Setup for Part 1
4	Failed	Setup for Part 2
5	Working	Setting up for Part 1
6	Working	Setting up for Part 2
7	Working	Setup for Part 1
8	Working	Setup for Part 2

Table 5.2: Modal Combinations for Subspaces

First Aggregation (Level 0 → Level 1)

The first step in the aggregation procedure is to separate the state space into ergodic and transient classes. In our case, this must be done for each individual subspace. Before we proceed with the separation, we define some terminology. An *almost transient* state is a state which is transient for ϵ=0, but not transient for $\epsilon \neq$0. When we refer to a *recurrent state*, we are referring to those states which are non-transient for both ϵ=0 and $\epsilon \neq$0.

When we separate the states into recurrent and (almost) transient classes there is more than one ergodic class for some of the subspaces. Table 5.3 lists the recurrent states for each subspace and numbers them according to ergodic class. In addition, the transient states associated with each subspace are listed (forming a transient class for each of subspaces 3 through 8). Recall that an ergodic class is a set of states, E, such that there is a non-zero probability given the system is in state $x_1 \in E$, at time t, that the system will enter any state $x_i \in E$ in some additional time Δt. At this time scale, for ϵ=0, there are no transitions leaving the recurrent states and therefore each of those 24 states must correspond to an ergodic class.

To help explain Table 5.3, consider the fourth subspace that consists of states

Subspace	Recurrent States	Ergodic Classes	Transient States
1	1	1	-
2	2	2	-
3	4,5	3,4	3
4	7,8	5,6	6
5	10-12	7-9	9
6	14-16	10-12	13
7	20,22,24,25,27,28	13-18	17,18,19,21,23,26
8	32,34,36,37,39,40	19-24	29,30,31,33,35,38

Table 5.3: Almost Transient and Recurrent States of Each Subspace

6,7 and 8, where state 7 and 8 are recurrent. If we refer back to Table 3.3, we see that machine 1 is failed and machine 2 is set up to operate on part type 2 for each of these states. State 6, the transient state, corresponds to a decision being made for machine 2. The transience of the state is a result of the fact that decisions are made very quickly, causing transitions to states 7 or 8. State 7 corresponds to no operations being performed and 8 to an operation on part type 2. Since the time scale is very short compared to the mean operation time, there is a high probability that the system will remain in state 7 or state 8 for the duration of the observation interval.

Since each ergodic class consists of a single state, the elements of the ergodic probability matrix can have two possible values when ϵ=0, zero for the almost transient states and 1 for the recurrent states. We can proceed to calculate the ergodic probability matrix in terms of its submatrices. For the first two subspaces, we obtain

$$\underline{U}_{11}^{(0)}(0), \underline{U}_{22}^{(0)}(0) = [1] \qquad (5.17)$$

with each of state 1 and state 2 being an individual ergodic class. Recalling that $\left[\underline{U}^{(0)}(0)\right]_{iJ}$ is the ergodic probability of state i given ergodic class J, we obtain for the remaining submatrices

$$\underline{U}_{33}^{(0)}(0), \underline{U}_{44}^{(0)}(0) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^T \tag{5.18}$$

$$\underline{U}_{55}^{(0)}(0), \underline{U}_{66}^{(0)}(0) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}^T \tag{5.19}$$

$$\underline{U}_{77}^{(0)}(0), \underline{U}_{88}^{(0)}(0) = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}^T . \tag{5.20}$$

Consider for example our previous example of subspace 4, correponding to submatrix $\underline{U}_{44}^{(0)}(0)$. States 7 and 8 are ergodic classes 5 and 6 so that

$$\left[\underline{U}^{(0)}(0)\right]_{75} = \left[\underline{U}^{(0)}(0)\right]_{86} = 1. \tag{5.21}$$

The correponding elements of the submatrix are $\left[\underline{U}_{44}^{(0)}(0)\right]_{21}$ and $\left[\underline{U}_{44}^{(0)}(0)\right]_{32}$ which are equal to 1, in agreement with equation (5.21). We also note that all of the off-diagonal submatrices of $\underline{U}^{(0)}(0)$ are identically 0. This is because the ergodic probability of a state, given an ergodic class to which it does not belong must be 0.

The class membership matrices require the calculation of the probabilities of entering each ergodic class from each almost transient state. The result is the set of equations (5.22) to (5.28) which defines the class membership matrices for this level. Note that these matrices are represented as functions of ϵ, but the matrices in (5.22) to (5.28) do not contain the parameter ϵ. This is because the transition rates for transitions originating in almost transient states all have the same order of magnitude, O(1). Therefore, although these matrices are in general a function of ϵ, the leading order terms in this case are O(1) and hence ϵ does not appear.

The first two submatrices correspond to the membership matrices for ergodic classes 1 and 2. Since each of these ergodic classes consists of a single state, the elements are simply 1.

$$\underline{V}_{11}^{(0)}(\epsilon), \underline{V}_{22}^{(0)}(\epsilon) = [1] \tag{5.22}$$

For $\underline{V}_{33}^{(0)}(\epsilon)$ and $\underline{V}_{44}^{(0)}(\epsilon)$, the elements corresponding to recurrent states are just 1 for the ergodic class to which the state belongs. The elements corresponding to almost transient states, make up the first column of $\underline{V}_{33}^{(0)}(\epsilon)$ and $\underline{V}_{44}^{(0)}(\epsilon)$ (states 3 and 6). These elements represent the probability of entering each ergodic class from the almost transient states.

$$\underline{V}_{33}^{(0)}(\epsilon) = \begin{bmatrix} \frac{\lambda_{20}(0,3)}{\lambda_{20}(0,3)+\lambda_{21}(0,3)} & 1 & 0 \\ \frac{\lambda_{21}(0,3)}{\lambda_{20}(0,3)+\lambda_{21}(0,3)} & 0 & 1 \end{bmatrix} \tag{5.23}$$

$$\underline{V}_{44}^{(0)}(\epsilon) = \begin{bmatrix} \frac{\lambda_{20}(0,4)}{\lambda_{20}(0,4)+\lambda_{22}(0,4)} & 1 & 0 \\ \frac{\lambda_{22}(0,4)}{\lambda_{20}(0,4)+\lambda_{22}(0,4)} & 0 & 1 \end{bmatrix} \tag{5.24}$$

Submatrices $\underline{V}_{55}^{(0)}(\epsilon)$ and $\underline{V}_{66}^{(0)}(\epsilon)$ are similar except there are three recurrent states in subspaces 5 and 6.

$$\underline{V}_{55}^{(0)}(\epsilon) = \begin{bmatrix} \frac{\lambda_{10}(1,1)}{\sum_j \lambda_{1j}(1,1)} & 1 & 0 & 0 \\ \frac{\lambda_{11}(1,1)}{\sum_j \lambda_{1j}(1,1)} & 0 & 1 & 0 \\ \frac{\lambda_{12}(1,1)}{\sum_j \lambda_{1j}(1,1)} & 0 & 0 & 1 \end{bmatrix} \tag{5.25}$$

$$\underline{V}_{66}^{(0)}(\epsilon) = \begin{bmatrix} \frac{\lambda_{10}(1,2)}{\sum_j \lambda_{1j}(1,2)} & 1 & 0 & 0 \\ \frac{\lambda_{11}(1,2)}{\sum_j \lambda_{1j}(1,2)} & 0 & 1 & 0 \\ \frac{\lambda_{12}(1,2)}{\sum_j \lambda_{1j}(1,2)} & 0 & 0 & 1 \end{bmatrix} \tag{5.26}$$

Finally, submatrices $\underline{V}_{77}^{(0)}(\epsilon)$ and $\underline{V}_{88}^{(0)}(\epsilon)$ are the most complicated because subspaces 7 and 8 contain six recurrent and six almost transient states. Each matrix contains six columns with entries of 1 or 0 corresponding to the recurrent states and six columns with fractional entries for the almost transient states, yielding

$\underline{V}_{77}{}^{(0)}(\epsilon) =$

$$\begin{bmatrix}
\frac{\lambda_{10}\lambda_{20}}{(\lambda_{21}+\lambda_{20})\sum_j \lambda_{1j}} & \frac{\lambda_{10}}{\sum_j \lambda_{1j}} & \frac{\lambda_{20}}{\lambda_{20}+\lambda_{21}} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{\lambda_{10}\lambda_{21}}{(\lambda_{21}+\lambda_{20})\sum_j \lambda_{1j}} & 0 & \frac{\lambda_{22}}{\lambda_{20}+\lambda_{21}} & 0 & \frac{\lambda_{10}}{\sum_j \lambda_{1j}} & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{\lambda_{11}\lambda_{20}}{(\lambda_{21}+\lambda_{20})\sum_j \lambda_{1j}} & \frac{\lambda_{11}}{\sum_j \lambda_{1j}} & 0 & 0 & 0 & 0 & \frac{\lambda_{20}}{\lambda_{20}+\lambda_{21}} & 1 & 0 & 0 & 0 & 0 \\
\frac{\lambda_{11}\lambda_{21}}{(\lambda_{21}+\lambda_{20})\sum_j \lambda_{1j}} & 0 & 0 & 0 & \frac{\lambda_{11}}{\sum_j \lambda_{1j}} & 0 & \frac{\lambda_{21}}{\lambda_{20}+\lambda_{21}} & 0 & 1 & 0 & 0 & 0 \\
\frac{\lambda_{12}\lambda_{20}}{(\lambda_{21}+\lambda_{20})\sum_j \lambda_{1j}} & \frac{\lambda_{12}}{\sum_j \lambda_{1j}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\lambda_{20}}{\lambda_{20}+\lambda_{21}} & 1 & 0 \\
\frac{\lambda_{12}\lambda_{21}}{(\lambda_{21}+\lambda_{20})\sum_j \lambda_{1j}} & 0 & 0 & 0 & \frac{\lambda_{12}}{\sum_j \lambda_{1j}} & 0 & 0 & 0 & 0 & \frac{\lambda_{21}}{\lambda_{20}+\lambda_{21}} & 0 & 1
\end{bmatrix} \tag{5.27}$$

In (5.27), the symbol λ_{ij} is used as a short form for $\lambda_{ij}(1,3)$. Similarly,

$\underline{V}_{88}{}^{(0)}(\epsilon) =$

$$\begin{bmatrix}
\frac{\lambda_{10}\lambda_{20}}{(\lambda_{22}+\lambda_{20})\sum_j \lambda_{1j}} & \frac{\lambda_{10}}{\sum_j \lambda_{1j}} & \frac{\lambda_{20}}{\lambda_{20}+\lambda_{22}} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{\lambda_{10}\lambda_{21}}{(\lambda_{22}+\lambda_{20})\sum_j \lambda_{1j}} & 0 & \frac{\lambda_{22}}{\lambda_{20}+\lambda_{22}} & 0 & \frac{\lambda_{10}}{\sum_j \lambda_{1j}} & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{\lambda_{11}\lambda_{20}}{(\lambda_{22}+\lambda_{20})\sum_j \lambda_{1j}} & \frac{\lambda_{11}}{\sum_j \lambda_{1j}} & 0 & 0 & 0 & 0 & \frac{\lambda_{20}}{\lambda_{20}+\lambda_{22}} & 1 & 0 & 0 & 0 & 0 \\
\frac{\lambda_{11}\lambda_{22}}{(\lambda_{22}+\lambda_{20})\sum_j \lambda_{1j}} & 0 & 0 & 0 & \frac{\lambda_{11}}{\sum_j \lambda_{1j}} & 0 & \frac{\lambda_{22}}{\lambda_{20}+\lambda_{22}} & 0 & 1 & 0 & 0 & 0 \\
\frac{\lambda_{12}\lambda_{20}}{(\lambda_{22}+\lambda_{20})\sum_j \lambda_{1j}} & \frac{\lambda_{12}}{\sum_j \lambda_{1j}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\lambda_{20}}{\lambda_{20}+\lambda_{22}} & 1 & 0 \\
\frac{\lambda_{12}\lambda_{22}}{(\lambda_{22}+\lambda_{20})\sum_j \lambda_{1j}} & 0 & 0 & 0 & \frac{\lambda_{12}}{\sum_j \lambda_{1j}} & 0 & 0 & 0 & 0 & \frac{\lambda_{22}}{\lambda_{20}+\lambda_{22}} & 0 & 1
\end{bmatrix} \tag{5.28}$$

where λ_{ij} represents $\lambda_{ij}(1,4)$ in this case. The transition matrix describing the aggregated system at the next slowest time scale can now be calculated using the formula given in Section 4.A. The non-zero submatrices of $\underline{A}^{(1)}(\epsilon)$ are given by equations (5.29) to (5.43). All submatrices that are not defined by these equations are identically zero. Note that the elements denoted by an asterisk have numerical values such that the elements in the corresponding column of the matrix sum to 0. The diagonal submatrices of the transition rate matrix for the model at time scale 1 are given by

$$\underline{A}_{11}^{(1)}(\epsilon), \underline{A}_{22}^{(1)}(\epsilon) = [*] \tag{5.29}$$

$$\underline{A}_{33}^{(1)}(\epsilon) = \begin{bmatrix} * & \frac{\lambda_{20}(0,3)\tau_{21}{}^{-1}}{\sum_{i=0,1}\lambda_{2i}(0,3)} \\ \frac{\lambda_{21}(0,3)\tau_{20}{}^{-1}}{\sum_{i=0,1}\lambda_{2i}(0,3)} & * \end{bmatrix} \tag{5.30}$$

$$\underline{A}_{44}^{(1)}(\epsilon) = \begin{bmatrix} * & \frac{\lambda_{20}(0,4)\tau_{22}{}^{-1}}{\sum_{i=0,2}\lambda_{2i}(0,4)} \\ \frac{\lambda_{22}(0,4)\tau_{20}{}^{-1}}{\sum_{i=2,0}\lambda_{2i}(0,4)} & * \end{bmatrix} \tag{5.31}$$

$$\underline{A}_{55}^{(1)}(\epsilon), \underline{A}_{66}^{(1)}(\epsilon) = \begin{bmatrix} * & \lambda_{10}^{(1)}\tau_{11}{}^{-1} & \lambda_{10}^{(1)}\tau_{12}{}^{-1} \\ \lambda_{11}^{(1)}\tau_{10}{}^{-1} & * & \lambda_{11}^{(1)}\tau_{12}{}^{-1} \\ \lambda_{12}^{(1)}\tau_{10}{}^{-1} & \lambda_{12}^{(1)}\tau_{11}{}^{-1} & * \end{bmatrix} \tag{5.32}$$

where

$$\lambda_{ij}^{(1)} = \frac{\lambda_{ij}(1,1)}{\sum_k \lambda_{ik}(1,1)} \text{ in } \underline{A}_{55}{}^{(1)}(\epsilon) \text{ while } \lambda_{ij}^{(1)} = \frac{\lambda_{ij}(1,2)}{\sum_k \lambda_{ik}(1,2)} \text{ in } \underline{A}_{66}{}^{(1)}(\epsilon).$$

$$\underline{A}_{77}^{(4)}(\epsilon) = \begin{bmatrix} * & c_{210} & c_{110} & 0 & c_{120} & 0 \\ c_{201} & * & 0 & c_{110} & 0 & c_{120} \\ c_{101} & 0 & * & c_{210} & c_{121} & 0 \\ 0 & c_{101} & c_{201} & * & 0 & c_{121} \\ c_{102} & 0 & c_{112} & 0 & * & c_{210} \\ 0 & c_{102} & 0 & c_{112} & c_{201} & * \end{bmatrix} \tag{5.33}$$

and

$$\underline{A}_{88}^{(1)}(\epsilon) = \begin{bmatrix} * & c_{220} & c_{110} & 0 & c_{120} & 0 \\ c_{202} & * & 0 & c_{110} & 0 & c_{120} \\ c_{101} & 0 & * & c_{220} & c_{121} & 0 \\ 0 & c_{101} & c_{202} & * & 0 & c_{121} \\ c_{102} & 0 & c_{112} & 0 & * & c_{220} \\ 0 & c_{102} & 0 & c_{112} & c_{202} & * \end{bmatrix} \tag{5.34}$$

where $c_{ijk} = \frac{\tau_{ij}{}^{-1}\lambda_{ik}(1,3)}{\sum_l \lambda_{il}(1,3)}$ in $\underline{A}_{77}^{(1)}(\epsilon)$ and $\frac{\tau_{ij}{}^{-1}\lambda_{ik}(1,4)}{\sum_l \lambda_{il}(1,4)}$ in $\underline{A}_{88}{}^{(1)}(\epsilon)$.

The off-diagonal submatrices, corresponding to transitions between the subspaces are given by

$$\underline{A}_{14}^{(4)}(\epsilon) = \begin{bmatrix} \epsilon f_s(0,4) & \epsilon f_s(0,4) \end{bmatrix}, \tag{5.35}$$

$$\underline{A}_{23}^{(1)}(\epsilon) = \begin{bmatrix} \epsilon f_s(0,3) & \epsilon f_s(0,3) \end{bmatrix}, \tag{5.36}$$

$$\text{and } \underline{A}_{58}^{(1)}(\epsilon), \underline{A}_{67}^{(1)}(\epsilon) = \begin{bmatrix} f & f & 0 & 0 & 0 & 0 \\ 0 & 0 & f & f & 0 & 0 \\ 0 & 0 & 0 & 0 & f & f \end{bmatrix}. \tag{5.37}$$

for the setup initiation events, where the symbol f represents $\epsilon f_s(1,4)$ in $\underline{A}_{58}^{(1)}$ and $\epsilon f_s(1,3)$ in $\underline{A}_{67}^{(1)}(\epsilon)$. For example, the (1,2) element of $\underline{A}_{14}^{(1)}(\epsilon)$ represents the transition rate from the second state of the fourth subspace to the first element of the first subspace. The transitions correspond to the initiation of setups for part type 1. The rates corresponding to setup completions in this model are given by

$$\underline{A}_{31}^{(1)}(\epsilon), \underline{A}_{42}^{(1)}(\epsilon) = \begin{bmatrix} \epsilon s^{-1} \\ 0 \end{bmatrix}, \tag{5.38}$$

$$\text{and } \underline{A}_{75}^{(1)}(\epsilon), \underline{A}_{86}^{(1)}(\epsilon) = \begin{bmatrix} \epsilon s^{-1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \epsilon s^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \epsilon s^{-1} \end{bmatrix}^T . \tag{5.39}$$

The failure events have rates in this model which are represented by

$$\underline{A}_{15}^{(1)}(\epsilon), \underline{A}_{26}^{(1)}(\epsilon) = \begin{bmatrix} 0 & \epsilon^2 p & \epsilon^2 p \end{bmatrix}, \tag{5.40}$$

$$\text{and } \underline{A}_{37}^{(1)}(\epsilon), \underline{A}_{48}^{(1)}(\epsilon) = \begin{bmatrix} 0 & 0 & \epsilon^2 p & 0 & \epsilon^2 p & 0 \\ 0 & 0 & 0 & \epsilon^2 p & 0 & \epsilon^2 p \end{bmatrix}, \tag{5.41}$$

while the repair rates appear in

$$\underline{A}_{51}^{(1)}(\epsilon), \underline{A}_{62}^{(1)}(\epsilon) = \begin{bmatrix} \epsilon^2 r \\ 0 \\ 0 \end{bmatrix}, \tag{5.42}$$

$$\text{and } \underline{A}_{73}^{(1)}(\epsilon), \underline{A}_{84}^{(1)}(\epsilon) = \begin{bmatrix} \epsilon^2 r & 0 & 0 & 0 & 0 & 0 \\ 0 & \epsilon^2 r & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{5.43}$$

These matrices completely define $\underline{A}^{(1)}(\epsilon)$, the transition rate matrix at time scale 1. As for the model at the original time scale, the submatrices not defined by equations (5.29) to (5.43) are identically 0 and represent one step transitions which are impossible. This new model at time scale 1 can be used to generate the model at time scale 2.

Second Aggregation (Level 1 → Level 2)

Examining the transition rates between subspaces, which are represented by the elements of the matrices $\underline{A}_{ij}^{(1)}(\epsilon)$ $i \neq j$, we note that they are at most O(ϵ) at this time scale. Therefore we can use the same partitioning of the state space for this aggregation as for the first aggregation. Examining the matrices $\underline{A}_{ii}(\epsilon)$ we see that they contain only O(1) elements. Since these elements represent the transition rates within each subspace, we see that each subspace has exactly one ergodic class. Solving for the ergodic probabilities we can generate the diagonal submatrices of the ergodic probability matrix $\underline{U}^{(1)}(0)$. Hence we obtain

$$\underline{U}^{(1)}_{11}(0), \underline{U}^{(1)}_{22}(0) = [0], \tag{5.44}$$

$$\underline{U}^{(1)}_{33}(0) = \begin{bmatrix} \frac{\lambda_{20}(0,3)\tau_{20}}{\lambda_{20}(0,3)\tau_{20}+\lambda_{21}(0,3)\tau_{21}} \\ \frac{\lambda_{21}(0,3)\tau_{21}}{\lambda_{20}(0,3)\tau_{20}+\lambda_{21}(0,3)\tau_{21}} \end{bmatrix}, \tag{5.45}$$

$$\underline{U}^{(1)}_{44}(0) = \begin{bmatrix} \frac{\lambda_{20}(0,4)\tau_{20}}{\lambda_{20}(0,4)\tau_{20}+\lambda_{22}(0,4)\tau_{22}} \\ \frac{\lambda_{22}(0,4)\tau_{22}}{\lambda_{20}(0,4)\tau_{20}+\lambda_{22}(0,4)\tau_{22}} \end{bmatrix}, \tag{5.46}$$

$$\underline{U}^{(1)}_{55}(0) = \begin{bmatrix} \frac{\lambda_{10}(1,1)\tau_{10}}{\sum_{k=0,1,2}\lambda_{1k}(1,1)\tau_{1k}} \\ \frac{\lambda_{11}(1,1)\tau_{11}}{\sum_{k=0,1,2}\lambda_{1k}(1,1)\tau_{1k}} \\ \frac{\lambda_{12}(1,1)\tau_{12}}{\sum_{k=0,1,2}\lambda_{1k}(1,1)\tau_{1k}} \end{bmatrix}, \tag{5.47}$$

$$\underline{U}^{(1)}_{66}(0) = \begin{bmatrix} \frac{\lambda_{10}(1,2)\tau_{10}}{\sum_{k=0,1,2}\lambda_{1k}(1,2)\tau_{1k}} \\ \frac{\lambda_{11}(1,2)\tau_{11}}{\sum_{k=0,1,2}\lambda_{1k}(1,2)\tau_{1k}} \\ \frac{\lambda_{12}(1,2)\tau_{12}}{\sum_{k=0,1,2}\lambda_{1k}(1,2)\tau_{1k}} \end{bmatrix}, \tag{5.48}$$

$$\underline{U}^{(1)}_{77}(0) = \begin{bmatrix} c_{00}(1,3) \\ c_{01}(1,3) \\ c_{10}(1,3) \\ c_{11}(1,3) \\ c_{20}(1,3) \\ c_{21}(1,3) \end{bmatrix} \tag{5.49}$$

$$\text{and } \underline{U}_{88}^{(1)}(0) = \begin{bmatrix} c_{00}(1,4) \\ c_{02}(1,4) \\ c_{10}(1,4) \\ c_{12}(1,4) \\ c_{20}(1,4) \\ c_{22}(1,4) \end{bmatrix} \tag{5.50}$$

where we have defined $c_{ij}(1,j)$ as

$$c_{ij}(1,j) = \frac{\lambda_{1i}(1,j)\tau_{1i}\lambda_{2j}(1,j)\tau_{2j}}{\left(\sum_{k=0,1,2}\lambda_{1k}(1,j)\tau_{1k}\right)\left(\sum_{l=0,j}\lambda_{2l}(1,j)\tau_{2l}\right)}. \tag{5.51}$$

The off-diagonal submatrices are again 0, because they correspond to the ergodic probabilities of states being in ergodic classes to which they do not belong.

The class membership matrix can be determined again by finding the diagonal submatrices which are given by

$$\underline{V}_{11}^{(1)}(\epsilon), \underline{V}_{22}^{(1)}(\epsilon) = [1], \tag{5.52}$$

$$\underline{V}_{33}^{(1)}(\epsilon), \underline{V}_{44}^{(1)}(\epsilon) = [1\ 1], \tag{5.53}$$

$$\underline{V}_{55}^{(1)}(\epsilon), \underline{V}_{66}^{(1)}(\epsilon) = [1\ 1\ 1], \tag{5.54}$$

$$\text{and } \underline{V}_{77}^{(1)}(\epsilon), \underline{V}_{88}^{(1)}(\epsilon) = [1\ 1\ 1\ 1\ 1\ 1]. \tag{5.55}$$

The off-diagonal matrices are identically zero, because the states are not members of more than one ergodic class (there are no transient or almost transient states).

Using the formula of Section 4.A, we can calculate the transition matrix for the model at level 2, which is given by

$$\underline{A}^{(2)}(\epsilon) = \begin{bmatrix} * & 0 & 0 & f_s(0,4) & \epsilon p^{(2)}(1,1) & 0 & 0 & 0 \\ 0 & * & f_s(0,3) & 0 & 0 & \epsilon p^{(2)}(1,2) & 0 & 0 \\ s^{-1} & 0 & * & 0 & 0 & 0 & \epsilon p^{(2)}(1,3) & 0 \\ 0 & s^{-1} & 0 & * & 0 & 0 & 0 & \epsilon p^{(2)}(1,4) \\ \epsilon r & 0 & 0 & 0 & * & 0 & 0 & f_s(1,4) \\ 0 & \epsilon r & 0 & 0 & 0 & * & f_s(1,3) & 0 \\ 0 & 0 & \epsilon r & 0 & s^{-1} & 0 & * & 0 \\ 0 & 0 & 0 & \epsilon r & 0 & s^{-1} & 0 & * \end{bmatrix}, \tag{5.56}$$

where

$$p^{(2)}(1,i) = p\left(\frac{\lambda_{11}(1,i)\tau_{11} + \lambda_{12}(1,i)\tau_{12}}{\sum_k \lambda_{1k}(1,i)\tau_{1k}}\right) \tag{5.57}$$

has been used to simplify the notation. We note that with only 8 states in the model at time scale 2 we are able to easily write out the transition rate matrix $\underline{A}^{(2)}(\epsilon)$ completely. The quantities $p^{(2)}(1,i)$ are equal to the failure rate multiplied by an expression which equals the fraction of time that machine 1 spends operating on parts. This expression is generated because the model assumes failures only occur when the machine is operating.

Third Aggregation (Level 2 → Level 3)

We note that since the number of states in the model at level 2 is 8, we were able to write out the transition matrix in equation (5.56). Therefore the calculations for the remaining aggregations can be written out completely, so state partitioning will not be used. To determine the appropriate model at time scale 3, we find the ergodic probability matrix which is given by

$$\underline{U}^{(2)}(0) = \begin{bmatrix} a_{10} & a_{20} & a_{30} & a_{40} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{11} & a_{21} & a_{31} & a_{41} \end{bmatrix}^T \tag{5.58}$$

where

$$\begin{aligned} a_{1i}, a_{2i} &= \frac{s f_s(i,3) f_s(i,4)}{2s f_s(i,3) f_s(i,4) \ + \ f_s(i,3) \ + \ f_s(i,4)}, & (5.59) \\ a_{3i} &= \frac{f_s(i,4)}{2s f_s(i,3) f_s(i,4) \ + \ f_s(i,3) \ + \ f_s(i,4)}, & (5.60) \\ \text{and } a_{4i} &= \frac{f_s(i,3)}{2s f_s(i,3) f_s(i,4) \ + \ f_s(i,3) \ + \ f_s(i,4)}. & (5.61) \end{aligned}$$

The expressions a_{ji} represent the fraction of time spent in set-up mode j given failure mode i. The class membership matrix is given by

$$\underline{V}^{(2)}(0) = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}. \tag{5.62}$$

Using $\underline{U}^{(2)}(0)$, $\underline{V}^{(2)}(0)$ and $\underline{A}^{(2)}(\epsilon)$, we can calculate the transition rate matrix for the model at next time scale. The result is

$$\underline{A}^{(3)}(\epsilon) = \begin{bmatrix} * & p^{(3)} \\ r & * \end{bmatrix} \tag{5.63}$$

where

$$p^{(3)} = \sum_{i=1}^{4} a_{i1} \ p^{(2)}(1,i) \tag{5.64}$$

is the failure rate in the model at time scale 3. This rate is the weighted sum of the failure rates in each set-up mode, with weights equal to the fraction of time the system spends in each mode.

Fourth Aggregation (Level 3 → Level 4)

At time scale 3, there are no ϵ terms in the transition rate matrix, but we proceed with the calculations to generate the information required for the calculation of expected event frequencies. The ergodic probability and class membership matrices are given by

$$\begin{aligned} \underline{U}^{(3)}(0) &= \begin{bmatrix} \frac{p^{(3)}}{p^{(3)}+r} \\ \frac{r}{p^{(3)}+r} \end{bmatrix} && (5.65) \\ \text{and } \underline{V}^{(3)}(\epsilon) &= [1\ 1]. && (5.66) \end{aligned}$$

Using equation (4.16) we find that the transition matrix at the final time scale is

$$\underline{A}^{(4)}(\epsilon) = [0]. \qquad (5.67)$$

Equation (5.67) indicates that the model of the system at this longest time scale consists of a single aggregate state. This is because the time intervals of observation are long enough that all of the events that are modeled become blurred and therefore cannot be distinguished individually.

5.A.2 Calculations for Production Frequencies.

This appendix uses the results of the time scale decomposition in Section 5.A.1 to calculate the expected frequencies of the events of interest in the system, namely part completions. The numerical results generated by the algorithm of Section 4.5.3 are provided here. A discussion of the results is presented in Sections 5.4 and 5.6. Since the time scale decomposition of the chain has already been performed in Section 5.A.1, we start by forming $\underline{Q}^{(0)}$ as described in step 2 of the algorithm. We define 4 sets of events, corresponding to the two parts produced and the two machines that can produce them. Each combination is assigned to a row of the matrix $\underline{Q}^{(0)}$ according to Table 5.4.

Row	Machine	Part
1	1	1
2	1	2
3	2	1
4	2	2

Table 5.4: Rows of the Event Frequency Matrix

The transitions that we count will be those which originate from states in which a machine of interest is operating on a part of interest and end in a state for which the machine is idle but capable of operating on a part. From Table 3.4, we see this is possible only for the completion of an operational activity on a machine, i.e. the production of a part. In this way, we obtain (5.68) as the matrix of expected frequencies at level 0.

$$\underline{Q}^{(0)} = \epsilon \left[\begin{array}{cccccccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \tau_{11}{}^{-1} & 0 & 0 & 0 & \tau_{11}{}^{-1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \tau_{12}{}^{-1} & 0 & 0 & 0 & \tau_{12}{}^{-1} \\ 0 & 0 & 0 & 0 & \tau_{21}{}^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \tau_{22}{}^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right. \cdots$$

$$\cdots \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & \tau_{11}{}^{-1} & \tau_{11}{}^{-1} & \tau_{11}{}^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \tau_{12}{}^{-1} & \tau^{12-1} & \tau_{12}{}^{-1} \\ 0 & 0 & 0 & 0 & \tau_{21}{}^{-1} & \tau_{21}{}^{-1} & 0 & 0 & \tau_{21}{}^{-1} & 0 & 0 & \tau_{21}{}^{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix} \cdots$$

$$\cdots \left. \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & \tau_{11}{}^{-1} & \tau_{11}{}^{-1} & \tau_{11}{}^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \tau_{12}{}^{-1} & \tau_{12}{}^{-1} & \tau_{12}{}^{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \tau_{22}{}^{-1} & \tau_{22}{}^{-1} & 0 & 0 & \tau_{22}{}^{-1} & 0 & 0 & \tau_{22}{}^{-1} \end{matrix} \right] \qquad (5.68)$$

Now, using the ergodic probability matrices from Section 5.A and the algorithm of Section 4.5.3 we obtain the formulae for $\underline{Q}^{(1)}$ through $\underline{Q}^{(4)}$ which are given by equations (5.69) through (5.72). The transition frequency matrix at time scale 1 is given by

$$\underline{Q}^{(1)} = \epsilon \left[\begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & \tau_{11}{}^{-1} & 0 & 0 & \tau_{11}{}^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \tau_{12}{}^{-1} & 0 & 0 & \tau_{12}{}^{-1} & 0 & 0 \\ 0 & 0 & 0 & \tau_{21}{}^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \tau_{21}{}^{-1} \\ 0 & 0 & 0 & 0 & 0 & \tau_{22}{}^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix} \right. \cdots$$

$$\cdots \left. \begin{matrix} \tau_{11}{}^{-1} & \tau_{11}{}^{-1} & 0 & 0 & 0 & 0 & \tau_{11}{}^{-1} & \tau_{11}{}^{-1} & 0 & 0 \\ 0 & 0 & \tau_{12}{}^{-1} & \tau_{12}{}^{-1} & 0 & 0 & 0 & 0 & \tau_{12}{}^{-1} & \tau_{12}{}^{-1} \\ 0 & \tau_{21}{}^{-1} & 0 & \tau_{21}{}^{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \tau_{22}{}^{-1} & 0 & \tau_{22}{}^{-1} & 0 & \tau_{22}{}^{-1} \end{matrix} \right] \qquad (5.69)$$

This matrix contains only the Markovian rates for transitions corresponding to operation completions. Proceding to time scale 2 we obtain

$$\underline{Q}^{(2)} = \epsilon \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\lambda_{21}(0,3)}{\sum_{k=0,1}\lambda_{2k}(0,3)\tau_{2k}} & 0 \\ 0 & 0 & 0 & \frac{\lambda_{22}(0,4)}{\sum_{k=0,2}\lambda_{2k}(0,4)\tau_{2k}} \\ \frac{\lambda_{11}(1,1)}{\sum_{k=0}^{2}\lambda_{1k}(1,1)\tau_{1k}} & \frac{\lambda_{12}(1,1)}{\sum_{k=0}^{2}\lambda_{1k}(1,1)\tau_{1k}} & 0 & 0 \\ \frac{\lambda_{11}(1,2)}{\sum_{k=0}^{2}\lambda_{1k}(1,2)\tau_{1k}} & \frac{\lambda_{12}(1,2)}{\sum_{k=0}^{2}\lambda_{1k}(1,2)\tau_{1k}} & 0 & 0 \\ \frac{\lambda_{11}(1,3)}{\sum_{k=0}^{2}\lambda_{1k}(1,3)\tau_{1k}} & \frac{\lambda_{12}(1,3)}{\sum_{k=0}^{2}\lambda_{1k}(1,3)\tau_{1k}} & \frac{\lambda_{21}(1,3)}{\sum_{k=0,1}\lambda_{2k}(1,3)\tau_{2k}} & 0 \\ \frac{\lambda_{11}(1,4)}{\sum_{k=0}^{2}\lambda_{1k}(1,4)\tau_{1k}} & \frac{\lambda_{12}(1,4)}{\sum_{k=0}^{2}\lambda_{1k}(1,4)\tau_{1k}} & 0 & \frac{\lambda_{22}(1,4)}{\sum_{k=0,2}\lambda_{2k}(1,4)\tau_{2k}} \end{bmatrix}^{T}. \tag{5.70}$$

This matrix contains the expected frequency of part completions for each part type, machine and aggregate state at time scale 2. These aggregate states correspond to each unique combination of failure and set-up modes. At time scale 3, we obtain a matrix of frequencies with only two columns:

$$\underline{Q}^{(3)} = \epsilon \begin{bmatrix} 0 & \sum_{j=1}^{4}\left(\frac{\lambda_{11}(1,j)a_{j2}}{\sum_{k=0,1,2}\lambda_{1k}(1,j)\tau_{1k}}\right) \\ 0 & \sum_{j=1}^{4}\left(\frac{\lambda_{12}(1,j)a_{j2}}{\sum_{k=0,1,2}\lambda_{1k}(1,j)\tau_{1k}}\right) \\ \frac{\lambda_{21}(0,3)a_{30}}{\sum_{k=0,1}\lambda_{2k}(0,3)\tau_{2k}} & \frac{\lambda_{21}(1,3)a_{31}}{\sum_{k=0,1}\lambda_{2k}(1,3)\tau_{2k}} \\ \frac{\lambda_{22}(0,4)a_{40}}{\sum_{k=0,2}\lambda_{2k}(0,4)\tau_{2k}} & \frac{\lambda_{22}(1,4)a_{41}}{\sum_{k=0,2}\lambda_{2k}(1,4)\tau_{2k}} \end{bmatrix}. \tag{5.71}$$

Each of these two columns corresponds to a different aggregate state in the model at time scale 3. These aggregate states in turn represent the two failure modes of the system. Finally, at time scale 4, all transitions are blurred, so there is only a single aggregate state. The long run production frequencies obtained at this time scale for each part and each machine are given by

$$\underline{Q}^{(4)} = \epsilon \left[\begin{array}{c} \left(\frac{r}{p^{(3)}+r}\right) \sum_{j=1}^{4} \frac{\lambda_{11}(1,j)a_{j2}}{\left(\sum_{k=0,1,2} \lambda_{1k}(1,j)\tau_{1k}\right)} \\ \left(\frac{r}{p^{(3)}+r}\right) \sum_{j=1}^{4} \frac{\lambda_{12}(1,j)a_{j2}}{\left(\sum_{k=0,1,2} \lambda_{1k}(1,j)\tau_{1k}\right)} \\ \left(\frac{p^{(3)}\lambda_{21}(0,3)a_{30}}{\sum_{k=0,1} \lambda_{2k}(0,3)\tau_{2k}} + \frac{r\lambda_{21}(1,3)a_{31}}{\sum_{k=0,1} \lambda_{2k}(1,3)\tau_{2k}}\right) \left(\frac{1}{r+p^{(3)}}\right) \\ \left(\frac{p^{(3)}\lambda_{22}(0,4)a_{40}}{\sum_{k=0,2} \lambda_{2k}(0,4)\tau_{2k}} + \frac{r\lambda_{22}(1,4)a_{41}}{\sum_{k=0,2} \lambda_{2k}(1,4)\tau_{2k}}\right) \left(\frac{1}{r+p^{(3)}}\right) \end{array} \right]. \tag{5.72}$$

Chapter 6

Conclusions

6.1 Remarks

The work in Chapters 2 through 5 of this thesis have used a continuous time finite state Markov Chain to model a Flexible Manufacturing System. Chapter 2 introduces the FMS environment and describes some of the features we might expect for a typical system. This description is then used in Chapter 3 to develop a Markov chain model for a simple FMS. Chapter 4 then presents some techniques that can be used to simplify and analyze such a model while Chapter 5 actually uses these techniques on the model generated in Chapter 3. The resources in the system were defined and the characteristics of those resources were used to define the states of the chain. The events within the FMS were represented by state transitions with rates determined by physical quantities.

A decomposition of the Markov Chain model of the entire system into several lower order models describing the system at various time scales was demonstrated. The results of the time scale decomposition were shown to be in agreement with what one would intuitively expect for models of the system at each time scale. The reduction in the complexity of the models with which one must work is significant, going from a 40 state chain to a worst case of 12 states. Once the decomposition was complete, the results were used to calculate production frequencies and determine production capacities.

The analysis demonstrates the usefulness of the formulation in providing concise, compact results regarding the behavior of the system. The following section provides a short example to demonstrate the lumping techniques for exactly lumpable systems. The usefulness of a generalization of this technique to almost lumpable systems is then discussed. Finally, we summarize the contributions of this thesis and the possibilities for future research in Sections 6.3 and 6.4 respectively.

6.2 Lumping Example

In Chapter 4, aggregation and lumping were introduced as methods of reducing the order of models describing a system. The aggregation technique was applied to a simple example in Chapter 5, in which a multiple time scale decomposition was developed for our model. This decomposition was then used as the basis for the application of the event frequency techniques of Section 4.5. This particular example did not present opportunities for the application of lumping (since the machines did not exhibit identical behavior). Therefore, this section uses a different example to demonstrate the usefulness of the lumping technique. In particular, a trivial example is used to show the reduction in complexity that can be obtained. In addition, suggestions for when the technique will be most useful are presented.

Suppose we have an FMS in which there are three identical machines. Furthermore, assume that we only wish to model the failure behavior of the system. This assumption does not affect the flavor of the results, but eliminates much of the

State	Machine 1	Machine 2	Machine 3
1	Failed	Failed	Failed
2	Failed	Failed	Working
3	Failed	Working	Failed
4	Failed	Working	Working
5	Working	Failed	Failed
6	Working	Failed	Working
7	Working	Working	Failed
8	Working	Working	Working

Table 6.1: Components for each state

calculations so that the main concepts are not obscured.

Let the rate at which a failure occurs for each machine be P and the repair rate be R, where these rates have identical meanings to those quantities introduced in Chapter 3. We also maintain the assumptions of exponentially distributed transition times, again so that the important results are not obscured.

To define the state space of the system, we recognize that it can be constructed as the Cartesian product of three components, with each component determining whether a particular machine is in a failed or working condition. The product of the components yields 8 states with the failure components for each state listed in Table 6.1. Since the transition rate for the failure event is P, the transition rate from state 5 to state 1, for example, is P because this transition corresponds to the failure of machine 1. Similarly, the transition rate from state 4 to state 8 is R because the transition results from the repair of machine 1. However, the rate from state 3 to state 5 is zero, because such a transition corresponds to an exactly simultaneous failure of machine 2 and repair of machine 1 and such an event is modeled as being impossible.

The state transition diagram for this chain is shown in Figure 6.1. We can write

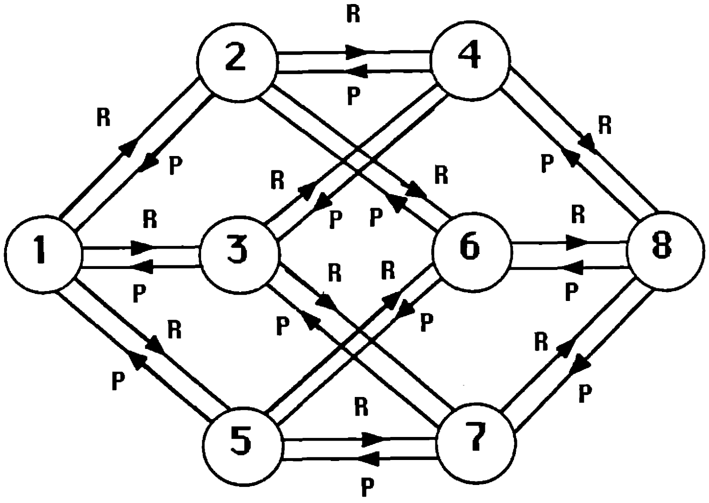


Figure 6.1: Exactly lumpable Markov chain

out the transition rate matrix for this chain obtaining

$$\underline{A} = \begin{bmatrix} * & P & P & 0 & P & 0 & 0 & 0 \\ R & * & 0 & P & 0 & P & 0 & 0 \\ R & 0 & * & P & 0 & 0 & P & 0 \\ 0 & R & R & * & 0 & 0 & 0 & P \\ R & 0 & 0 & 0 & * & P & P & 0 \\ 0 & R & 0 & 0 & R & * & 0 & P \\ 0 & 0 & R & 0 & R & 0 & * & P \\ 0 & 0 & 0 & R & 0 & R & R & * \end{bmatrix} . \tag{6.1}$$

For this example, one possible way of lumping the states would be to lump states 2,3 and 5 together as well as states 4,6 and 7. Referring to Table 6.1, we see that this corresponds to lumping all of the states which correspond to an equal number of machines working, regardless of which machines are working. For example, states 2,3 and 5 each represent states for which only 1 of the 3 machines is in working

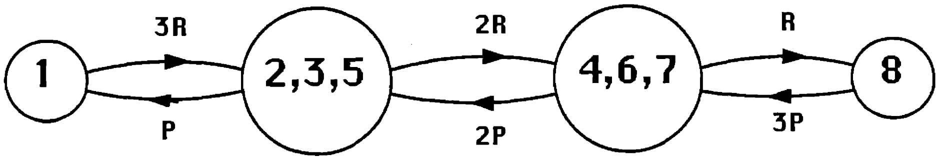


Figure 6.2: Lumped chain for 3 machine example

order. Therefore this lumping would be appropriate if our analysis only required knowledge of how many machines are working or failed.

Next, we perform the actual calculations. Referring to Section 4.4, we calculate $\underline{C}$, $\underline{B}$, and $\underline{A}_c$, the last of which is the lumped transition rate matrix and is given by

$$\underline{A}_c = \begin{bmatrix} * & P & 0 & 0 \\ 3R & * & 2P & 0 \\ 0 & 2R & * & 3P \\ 0 & 0 & R & * \end{bmatrix}. \tag{6.2}$$

The state transition diagram for the lumped chain is provided in Figure 6.2. Therefore we see that the state space has been reduced from 8 states to 4.

This reduction in the size of the state space is the principle reason for using the lumping technique. In this example, the procedure is somewhat academic because the original model had only 8 states without lumping. However, this small example was chosen only so that the important features of the calculations could be observed. The technique is easily extended to much larger models. For example, when modeling systems with ten machines, the failure dynamics alone yield 1024 states. If the results from our example are generalized to this 10 machine case, the lumped model has only 11 states.

Examining these results however, we note that a similar result would have been achieved if the model had been set up originally with the state defined by the number of machines that are in working order. However, if we consider a case where the machines are not exactly the same, then directly setting up the model in this way is not possible.

Note that if the state space is formed using a complete enumeration of which machines are failed, we may obtain a Markov chain that is *almost lumpable*. Almost lumpable refers to a chain whose dynamics depend on a parameter ϵ, and which is lumpable for $\epsilon = 0$, but not lumpable for $\epsilon \neq 0$. The theory regarding when and how such a system can be lumped with errors that uniformly approach 0 as $\epsilon \rightarrow 0$ has not been developed to date. Therefore, this is left as a possible future research topic.

6.3 Contributions of this thesis

The contributions provided by this thesis are as follows.

Using a finite state Markov chain model of an FMS that can be generalized for many activities and events, a time scale decomposition was effected using the algorithmic approach of Rohlicek. This technique will also be applicable to alternate models including additional aspects or details of an FMS. In addition, a physical interpretation was given corresponding to the numerical aggregation results, providing a qualitative demonstration of the results that can be expected when the aggregation techniques are employed.

A method for calculating the expected frequencies of transitions within a Markov chain was provided. The method, which essentially requires one to take a series of expected values, parallels the concepts in Gershwin [10], but provides a compact, precise means of performing the calculations. The use of the expressions obtained in these calculations to find capacity constraints as defined in [10] was also demonstrated.

The effects of almost transient states on the calculation of the expected frequency of transitions were described. In particular, it was demonstrated that calculations which used ergodic probabilities that are determined by setting the perturbation parameter ϵ, equal to zero, lead to incorrect results. In fact, this approach can yield relative errors of up to 100 percent for some expected frequencies. Therefore, a modification of the expected frequency calculations was developed which accounts for the leading order terms (as powers of ϵ) of the ergodic probabilties. The resulting technique retains the analytical simplicity of calculating ergodic probabilies with $\epsilon = 0$, but generates approximations of the expected frequencies of events that are asymptotically exact as $\epsilon \rightarrow 0$.

The issue of ergodicity and the implications for the expected frequencies that are calculated was also addressed. In particular, a set of conditions was provided for which the time average frequency of an event will equal the expected value of the frequency with probability 1 as $\epsilon \rightarrow 0$.

6.4 Future Research Directions

This thesis was intended to show how time scale decomposition techniques can be used to aid in the modeling Flexible Manufacturing Systems. Techniques that can be applied to the decomposed model to obtain production rates and capacities were derived and illustrated. Using this as a starting point, there are several research areas worthy of investigation.

Techniques for increasing the generality of this type of model could be investigated. In particular, the results of Rohlicek [23] handle a much broader class of models than the Markov Chains considered here. Therefore, the time scale decomposition techniques should apply to models which, for instance, do not require the assumption of exponentially distributed transition rates. This will allow more realistic models, particularly for the portions of the model describing quantities which are unlikely to be random. For example, it is likely that a realistic model would represent the holding times for the transitions corresponding to the completion of a decision as nearly deterministic quantities, since a set of decision rules should take a consistent amount of time to implement.

As mentioned in Section 6.2, the generalization of lumping techniques to systems which are almost lumpable would also be useful for obtaining tractable models for very large and complex manufacturing environments. The almost lumpable terminology refers to the case when certain physical quantities in the system, represented by state transition rates in the chain, are nearly identical, but differ by a quantity proportional to ϵ, a small positive number. Therefore, if $\underline{A}(\epsilon)$ is the transition matrix for a Markov Chain with almost lumpable states, we could write it as

$$\underline{A}(\epsilon) = \underline{A}(0) + \underline{B}(\epsilon), \tag{6.3}$$

where $\underline{A}(0)$ is the rate transition matrix for an exactly lumpable chain and $\underline{B}(\epsilon)$ is a matrix whose elements are at most $O(\epsilon)$. The ability to handle these cases would greatly increase the maximum complexity of systems that could be handled using the techniques of this paper. For example, we saw in Section 6.2 that for a 10 machine system the state space could be reduced from 1024 to 11 states if the machines are identical. If an almost lumpable approach were developed, a comparable reduction might be possible, even if the machines were similar, but not identical.

There are also many additional features of an FMS that could be incorporated into a model, and investigated in the multiple time scale framework. For example, this thesis considered the case where a part only required an operation by a single machine in the system. The case where parts visit a number of work stations in sequence could also be considered. The interaction of the two machines in terms of part flow could be modeled, possibly with various buffers of work in progress between them, and the resulting dynamics over various time scales analyzed.

Once the model for a Flexible Manufacturing System has been formed, the techniques of Chapter 4 can be applied to obtain the expected frequency of various events of interest; however, there are many additional uses that could be investigated for this type of model. In particular, the relationship between the expected frequencies at adjacent levels could be examined. Gershwin [9] discusses the concept of information flow up and down the hierarchy, and hence between levels. He introduces the relationship between this flow of information and the concepts of capacity and control in the FMS. An analysis of the expected frequencies at various levels can be used to obtain compact, general equations for capacity constraints as demonstrated in Section 6.3. An investigation into the use of the frequency results for the evaluation of control policies would also be of interest.

The issue of trying to optimize control policies for the Flexible Manufacturing System can also be addressed. Some attention has been given in the literature to the near optimal control of Markov Chains which exhibit behavior at multiple time scales. Approaches by Delebecque and Quadrat [7], Philips and Kokotovic [21], and Larson [19], deal with the determination of near optimal control policies for Markov Chains using reduced-order models at two time scales. The controls that are obtained by these schemes are nearly optimal in the sense that the value that is obtained for an objective function when the nearly optimal control policy is applied differs from the value obtained when the optimal policy is applied by an amount which is $O(\epsilon)$. It is the possibility of being able to find such a control policy that makes the time scale analysis particularly valuable from a practical point of view when an optimal solution to the control problem may not be possible.

Bibliography

[1] R. Akella., Y. Choong, and S.B. Gershwin. Performance of hierarchical production scheduling policy. *IEEE Transactions on Components, Hybrids and Manufacturing Technology*, 7(2), September 1984.

[2] M. Berrada and K.E. Stecke. A branch and bound approach for machine load balancing in flexible manufacturing systems. *Management Science*, 32(10), October 1986.

[3] J.A. Buzacott. Optimal operating rules for automated manufacturing systems. In *Proceedings of the* 19th *IEEE Conference on Decision and Control*, Albuquerque, New Mexico, 1980.

[4] J.A. Buzacott and J.G. Shanthikumar. On approximate queueing models of dynamic job shops. *Management Science*, 31(7), July 1985.

[5] M. Coderch, A.S. Willsky, S.S. Sastry, and D.A. Castanon. Hierarchical aggregation of linear systems with multiple time scales. *IEEE Transactions on Automatic Control*, 30(11), November 1983.

[6] P.J. Courtois. *Decomposability.* Academic Press, NY, 1977.

[7] F. Delebecque and J.P. Quadrat. Optimal control of markov chains admitting strong and weak interactions. *Automatica*, 17(2), March 1981.

[8] F. Delebecque, J.P. Quadrat, and P.V. Kokotovic. A unified view of aggregation and coherency in networks and markov chains. *International Journal of Control*, 40(5), November 1984.

[9] S.B. Gershwin. *A Hierarchical Framework for Discrete Event Scheduling in Manufacturing Systems.* Technical Report 1682, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, August 1987.

[10] S.B. Gershwin. *A Hierarchical Framework for Manufacturing Systems Scheduling.* Technical Report 1, Laboratory for Manufacturing and Productivity, Massachusetts Institute of Technology, September 1987.

[11] S.B. Gershwin. Stochastic scheduling and set-ups in flexible manufacturing systems. In *Proceedings of the Second ORSA/TIMS Conference on Flexible Manufacturing Systems: Operations Research Models and Applications*, Amsterdam, 1986.

[12] R.R. Hildebrant. Scheduling flexible machining systems. In *Proceedings of the* 19th *IEEE Conference on Decision and Control*, Albuquerque, New Mexico, 1980.

[13] Y.C. Ho and X. Cao. Perturbation analysis and optimization of queueing networks. *Journal of Optimization Theory and Applications*, 40(4), August 1983.

[14] Y.C. Ho and C. Cassandras. A new approach to the analysis of discrete event dynamic systems. *Automatica*, 19(4), August 1983.

[15] G.K. Hutchinson. The control of flexible manufacturing systems: required information and algorithm structures. In *IFAC Symposium on Information Control and Problems in Manufacturing Technology*, 1977.

[16] Kielson. *Markov Chain Models- Rarity and Exponentiality.* Springer-Verlag, 1979.

[17] J. Kimemia and S.B. Gershwin. An algorithm for the computer control of a flexible manufacturing system. *IIE Transactions*, September 1983.

[18] B.J. Lageweg, J.K. Lenstra, and A.H.G. Rinnooy Kan. Job shop scheduling by implicit enumeration. *Management Science*, 24(4), December 1977.

[19] R.E. Larson. *Research and Development of a Unified Approach to Operations Scheduling for Electric Power under Uncertainty.* Technical Report DOE/ET/29243-T1, Systems Control Inc., for the United States Department of Energy, 1985.

[20] G.J. Olsder and R. Suri. Time optimal control of parts-routing in a manufacturing system with failure prone machines. In *Proceedings of the* 19th *IEEE Conference on Decision and Control*, Albuquerque, New Mexico, 1980.

[21] R.G. Philips and P.V. Kokotovic. A singular perturbation approach to modeling and control of markov chains. *IEEE Transactions on Automatic Control*, 26(5), October 1981.

[22] D.A. Pinto, D.G. Dannenberg, and B.M. Khumawala. Assembly line balancing with processing alternatives: an application. *Management Science*, 29, July 1983.

[23] 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.

[24] S.M. Ross. *Introduction to Probability Models.* Academic Press Inc., 1980.

[25] K.E. Stecke. Formulation and solution of nonlinear integer problems for flexible manufacturing systems. *Management Science*, 29(3), March 1983.

[26] K.E. Stecke. Planning and control models to analyze problems of flexible manufacturing. In *Proceedings of the* 23rd *IEEE Conference on Decision and Control*, 1984.