Fluid passage-time calculation in large Markov models

Richard A. Hayden[a], Jeremy T. Bradley[a]

[a] *Dept. of Computing, Imperial College London,*
Huxley Building, 180 Queen's Gate, London SW7 2BZ, United Kingdom

Abstract

Recent developments in the analysis of large Markov models facilitate the fast approximation of transient characteristics of the underlying stochastic process. So-called *fluid analysis* makes it possible to consider previously intractable models whose underlying discrete state space grows exponentially as model components are added. In this work, we show how fluid approximation techniques may be used to extract passage-time measures from performance models. We focus on two types of passage measure: passage-times involving individual components; as well as passage-times which capture the time taken for a population of components to evolve.

Specifically, we show that for models of sufficient scale, passage-time distributions can be well approximated by a deterministic fluid-derived passage-time measure. Where models are not of sufficient scale, we are able to generate approximate bounds for the entire cumulative distribution function of these passage-time random variables, using moment-based techniques. Finally, we show that for some passage-time measures involving individual components the cumulative distribution function can be directly approximated by fluid techniques.

Key words: stochastic process algebra, fluid approximation, passage-time, numerical bounds

1. Introduction

Passage-time or *response-time* distributions are some of the most sought-after quantitative performance measures of a system. Passage-time quantiles form the basis of many service level agreements (SLAs) in the telecommunications and other industries, e.g. a broadband connection should be successfully established within 2 seconds, 95% of the time.

However, analysis of such industrial-scale systems requires the ability to deal with massive underlying discrete state spaces which grow exponentially as system components are added to the model. Indeed the capability of traditional explicit state-space techniques for computing passage-time distributions is quickly exceeded [1].

Fluid analysis of performance models offers the exciting potential for the analysis of massive state-spaces at small computational cost. We consider the case of massively parallel Markov models that consist of synchronising groups of component Markov chains. In this case, fluid analysis involves approximating the underlying discrete state space with continuous real-valued variables and describing the transient evolution of those variables with ordinary differential equations (ODEs). Specifically, the solution to the ODEs is an approximation to certain discrete stochastic processes which count the number of Markov chain components in the model which are in a given state.

It has previously been shown that many passage-time measures are equivalent to the *time-to-extinction* of a certain set of components within a modified model [2]. This is a quantity which, using fluid analysis techniques, can be approximated by the time it takes for a component of the system of ODEs to reach zero.

In Section 2, we build upon the notion of a *global passage time* as a means of capturing system-wide passages. We show how they can be approximated by fluid analysis as a time-to-extinction measure. We create a limit relationship for the global passage-time expression. In particular, we show that, in a limit of increasing model size, the CTMC passage time densities converge to a deterministic distribution which matches the deterministic approximation of the fluid technique (Section 3).

From this work, we confirm that a solely deterministic representation of a random variable passage time has shortcomings. Specifically, there is no longer a notion of probability distribution in an ODE solution for what is, after all, a stochastic quantity. This, is the motivation for the second contribution of the paper (Section 2.1.2 and 4.3), which provides efficient approximate upper and lower bounds on the cumulative distribution function of the entire passage-time.

Section 2 also shows how *individual passage times* can track the evolution of single components in massively parallel systems. For these individual passage times the entire cumulative distribution function can be well approximated by fluid analysis. In Section 3, we provide convergence proofs for both global and individual passage-times. Section 4 contains a client–server worked example which demonstrates and compares all the passage-time calculation techniques in operation.

In summary, we aim to provide a machinery for the systematic approximation of passage times in performance models with underlying state-space sizes well beyond the capabilities of existing techniques. In order to accomplish this we will use the stochastic process algebra, PEPA, to express the types of massively parallel system that we wish to analyse.

1.1. Introduction to PEPA

We begin by introducing PEPA [3, 4], which is a simple stochastic process algebra, but one which has sufficient expressiveness to model a wide variety of systems, including multimedia applications [5], mobile phone usage [6], GRID scheduling [7], production cell efficiency [8] and web-server clusters [9] amongst others. It is particularly adept at capturing large parallel software systems, such as peer-to-peer networks [10], to which the style of fluid analysis considered here is particularly suited.

As in all process algebras, systems are represented in PEPA as the composition of *components* which undertake *actions*. In PEPA the actions are assumed to have a duration. Thus the expression $(\alpha, r).P$ denotes a component which can undertake an α-action, at rate r to evolve into a component P. Here $\alpha \in \mathcal{A}$ where $\mathcal{A}$ is the set of action types and $P \in \mathcal{C}$ where $\mathcal{C}$ is the set of component types. The rate r is interpreted as a random delay sampled from an exponential distribution with parameter r. This means that the stochastic behaviour of the model is governed by an underlying continuous-time Markov chain, the explicit definition of which will be given later in this section.

PEPA has a small set of combinators, allowing system descriptions to be built up as the concurrent execution and interaction of simple sequential components. The syntax of the type of PEPA model considered in this paper may be specified formally using the grammar:

$$\begin{aligned} S &::= (\alpha, r).S \;\mid\; S + S \;\mid\; C_S \\ P &::= P \underset{L}{\bowtie} P \;\mid\; P/L \;\mid\; C \end{aligned} \tag{1.1}$$

where S represents a *sequential component* and P represents a *model component* which executes in parallel. C stands for a constant which denotes either a sequential component or a model component as introduced by a definition. C_S stands for constants which denote only sequential components. The effect of this syntactic separation between these types of constants is to constrain legal PEPA components to be cooperations of only sequential processes.

The structured operational semantics can be found in [3, Chap. 3]. A brief discussion of the basic PEPA operators is given below:

Prefix The basic mechanism for describing the behaviour of a system with a PEPA model is to give a component a designated first action using the prefix combinator, denoted by a full stop. As explained, $(\alpha, r).P$ carries out an α-action with rate r, and it subsequently behaves thereafter as P.

Choice The component $P + Q$ represents a system which may behave either as P or as Q. The activities of both P and Q are enabled. The first activity to complete selects the path to be executed in the model: the other possible path is discarded. The system then proceeds by taking on the behaviour of the derivative resulting from the evolution of the chosen component.

Constant It is convenient to be able to assign names to patterns of behaviour associated with components. Constants are components whose meaning is given by a defining equation. The notation for this is $X \stackrel{def}{=} E$. This also allows the recursive definition of components, for example, $X \stackrel{def}{=} (\alpha, r).X$ performs α at rate r forever.

Hiding The possibility to abstract away some aspects of the behaviour of a component is provided by the hiding operator, denoted P/L. Here, the set L identifies those activities which are to be considered internal or private to the component and which will appear as the hidden action type τ in the transition system of the model. We will not consider hiding further in this paper although the hiding translation of [11] to fluid model would work in this context.

Cooperation We write $P \underset{L}{\bowtie} Q$ to denote cooperation between P and Q over L. The set which is used as the subscript to the cooperation symbol, the *cooperation set* L, determines those activities on which the components are forced to synchronise. For action types not in L, the components proceed independently and concurrently with their enabled activities. We write $P \parallel Q$ as an abbreviation for $P \underset{\emptyset}{\bowtie} Q$, where P and Q execute in parallel.

Fundamental to PEPA is the notion of *apparent rate*, $r_\alpha(P)$, which measures the observed rate that a process, P, executes an action, α. This defines the rate that a cooperating process sees and is therefore integral to the speed of cooperation between processes. A formal definition can be found in Appendix A.1.

In process cooperation, if a component enables an activity whose action type is in the cooperation set it will not be able to proceed with that activity until the other component also enables an activity of that type. The two components then proceed together to complete the *shared activity*. Once enabled, the rate of a shared activity has to be altered to reflect the slower component in a cooperation. Within the cooperation framework, PEPA assumes *bounded capacity*: that is, a component cannot be made to perform an activity faster by cooperation, and the rate of a shared activity is defined as the minimum of the apparent rates of the activity in the cooperating components. This is discussed in more detail in [3].

In some modelling situations, the rate of a shared activity is determined by only a subset of components in a cooperation (the *active* partners). Other components may be *passive* partners in this cooperation. It is then only the *ability* of the passive partners to perform the shared action which is required for it to be able to proceed; the stochastic duration of the shared action is otherwise determined solely by the active partners. We do not consider passive cooperation in this paper since it is tricky to handle in a fluid analysis context, although we have a successful approach which could be applied to passage-time analysis as well [12].

1.1.1. Execution strategy

For a given PEPA component C, we define its *derivative set* $ds(C)$ as the set of components reachable from C. That is, $ds(C)$ is the smallest set of components such that $C \in ds(C)$ and if for any $C_1 \in ds(C)$, $C_1 \xrightarrow{(\alpha, r)} C_2$ then $C_2 \in ds(C)$.

For a given PEPA component C, we may then naturally construct its *derivation graph*, a labelled and directed multigraph. The nodes of this multigraph are the derivative states of C, that is, the set of nodes is $ds(C)$. Two nodes in the multigraph, say C_1 and $C_2 \in ds(C)$, have a directed arc between them for every

transition $C_1 \xrightarrow{(\alpha, \lambda)} C_2$. The label of this arc is then the activity corresponding to the transition, that is, (α, λ).

The derivation graph can then be interpreted naturally as a CTMC, whose states are given by the nodes (i.e. derivative states) and each arc represents a transition at the rate of the activity labelling the arc. We call this the *underlying CTMC* of the model. The full operational semantics of PEPA are presented in [3, Chap. 3].

1.2. A Motivating Example

We consider the ubiquitous situation of many processors running in parallel, but each in regular need of some resource (perhaps for example, communications channels or storage devices). The type of model we wish to consider in this paper is one which exhibits massive parallelism. We present such a system below, where a large number of parallel processors cooperate with a pool of parallel resources. In classical Markov chain analysis of any variety, this would require exploration of the global state space and, even for such a simple system, this quickly becomes computationally infeasible.

We capture the scenario of n processors cooperating on a $task_1$ action with m resources with the following PEPA system equation:

$$PR(n, m) \stackrel{def}{=} Processor_0[n] \underset{\{task_1\}}{\bowtie} Resource_0[m] \tag{1.2}$$

where $C[n]$ represents n parallel copies of component C:

$$C[n] := \underbrace{(C \parallel \ldots \parallel C)}_{n} \tag{1.3}$$

Each processor is represented as a $Processor_0$ component and each resource as a $Resource_0$ component. Each processor operates forever in a simple loop, completing two tasks in sequence, $task_1$ and then $task_2$:

$$Processor_0 \stackrel{def}{=} (task_1, r_1).Processor_1$$
$$Processor_1 \stackrel{def}{=} (task_2, q_1).Processor_0$$

The resources on the other hand first complete a $task_1$ action also, but then complete a *reset* action:

$$Resource_0 \stackrel{def}{=} (task_1, r_2).Resource_1$$
$$Resource_1 \stackrel{def}{=} (reset, q_2).Resource_0$$

The $task_1$ action is a shared action between the processors and resources to model the situation of a processor having to acquire a resource which it needs to complete its first task. The actions $task_2$ and *reset*, on the other hand, will not be shared, meaning they are completed independently and without synchronisation by the processors and resources respectively. The cooperation over $task_1$ is an instance of active cooperation and, for the simplest case of $n = m = 1$, completes at rate $\min(r_1, r_2)$ according to the operational semantics of PEPA [3].

Still in the case of $n = m = 1$, Figure 1 gives the underlying CTMC explicitly, adopting the shorthand P_i for $Processor_i$ and R_i for $Resource_i$.

The model $PR(n, m)$ has n processor components and m resource components, each of which can be in one of two states, so the underlying CTMC of this simple model has 2^{n+m} states, that is, exponential growth in the number of processors and resources. Such rapid growth in the size of the state space for models of only modest description is known as the *state space explosion problem*. It would of course be even more pronounced for more realistic and detailed models of distributed systems.

In this paper, we are going to extract passage time densities and distributions from such a model description without having to expand the global state space. We will show that, using fluid techniques, we can answer two types of passage-time question:

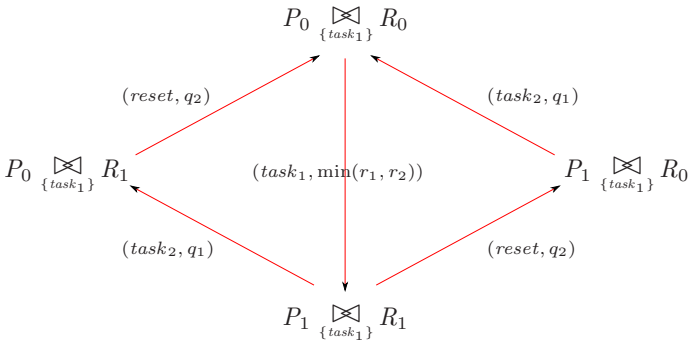


Fig. 1: Underlying CTMC for simple processor/resource model

Global passage time. What is the probability that half the processors have executed at least one $task_1$ action by time t.

Individual passage time. What is the probability that any individual resource has completed five *reset* actions by time t.

We will introduce these two classes of passage-times in Section 2. In Section 3 we will show, in the case of *global passage times*, that there is a passage-time limit relation that can be expressed for models such as $PR(2n, n)$ such that in the limit of $n \to \infty$ the sequence of passage-time densities will converge to a deterministic distribution. For models where n is not large enough for this to be an accurate approximation, we will also show that it is possible to estimate easily-calculated bounds on the CDF of the passage time, again using fluid techniques.

In the case of *individual passage times*, we will also show that fluid analysis can be used to approximate the cumulative distribution function of the passage-time measure directly.

1.3. Fluid Analysis of PEPA Models

Fluid analysis captures the number of components in a particular derivative state of a PEPA model as the system evolves. The evolution of the PEPA components is described by a set of ordinary differential equations, derived directly from the PEPA model description. These differential equations are easy to solve numerically and provide a straight-forward approach to analysing massive performance models. Fluid semantics for PEPA, first introduced by Hillston [13], have since been extended and developed in a number of different directions in the literature [14, 15, 11]. Furthermore, similar ideas have been applied in other stochastic process algebra [16, 17] and stochastic Petri net [18] formalisms.

From Hayden *et al.* [11], we conservatively extend the standard PEPA grammar of Equation (1.1) to support explicit identification of *component groups* using *component group labels*, defining the notion of a *grouped PEPA model*. These component groups will be used to identify the parallel component populations of the Markov model, and thus the level at which fluid analysis is performed.

1.3.1. Grouped PEPA models

We begin by defining a component group D, which is simply a parallel cooperation (involving no synchronisation) of standard PEPA components P:

$$D ::= D \parallel D \;\mid\; P \tag{1.4}$$

A *grouped PEPA model* M is then an arbitrary combination of labelled component groups:

$$M ::= M \underset{L}{\bowtie} M \;\mid\; Y\{D\} \tag{1.5}$$

where L is a set of action types. The term $Y\{D\}$ is a *labelled component group* and extends the original PEPA syntax. Y is a unique component group label drawn from some sufficiently large label set. Action hiding at the level of grouped PEPA models is discussed in detail in [11] and would equally apply here.

The operational semantics for this augmented version of PEPA are the natural extension of the standard PEPA operational semantics [3] and are given formally in [11]. The only difference is that the explicit identification of component groups is maintained as the model evolves. A *flattening function*, $\mathcal{F}(G)$, which yields the corresponding standard PEPA model by simply removing the component group labels from the grouped PEPA model G is defined formally in Appendix A.2.

The set of derivative states is defined similarly, but each derivative state also maintains its explicit component group labelling. For example, we might represent the model $PR(n,\, m)$ introduced earlier as the grouped PEPA model:

$$PR_G(n,\, m) \stackrel{def}{=} \mathbf{Processors}\{Processor_0[n]\} \underset{\{task_1\}}{\bowtie} \mathbf{Resources}\{Resource_0[m]\} \tag{1.6}$$

where the definition of the processor and resource components are as before. Note that $\mathcal{F}(PR_G(n,\, m)) = PR(n,\, m)$. That is, $PR(n,\, m)$ has exactly the same operational semantics (and thus underlying CTMC) as $PR_G(n,\, m)$, the only difference is that component groups are made explicit in the latter model.

As the model evolves, the component groupings are maintained, for example, one possible evolution and grouped derivative state of this model is:

$$\begin{aligned}\mathbf{Processors}\{P_0[n]\} \underset{\{task_1\}}{\bowtie} \mathbf{Resources}\{R_0[m]\} &\xrightarrow{(task_1,\, \min(nr_1,\, mr_2)/nm)} \\ \mathbf{Processors}\{P_1 \parallel P_0[n-1]\} &\underset{\{task_1\}}{\bowtie} \mathbf{Resources}\{R_1 \parallel R_0[m-1]\}\end{aligned} \tag{1.7}$$

The purpose of this simple syntactic extension to PEPA is to allow a much clearer presentation of the fluid semantics. In this particular case, the two component groups (identified by the labels **Processors** and **Resources**) specify that the fluid analysis will happen at the level of the P_0, P_1, R_0 and R_1 components. That is, these are the four derivative states we will count copies of; there will be one differential equation defined for each of these four component states.

We now define some key functions of a grouped PEPA model that will be used to generate the fluid model in Definition 1.4. The table below provides some short definitions; formal definitions can be found in Appendix A.3. In the examples below, we have adopted the further shorthand for the component group labels, **P** for **Processors** and **R** for **Resources**.

$\mathcal{G}(G)$	The set of all component group labels in the grouped PEPA model G, e.g. $\mathcal{G}(PR_G(n,\, m)) = \{\mathbf{P},\, \mathbf{R}\}$.
$\mathcal{B}(G,\, H)$	The set of all local standard PEPA component states in the component group of G which has group label H, e.g. $\mathcal{B}(PR_G(n,\, m),\, \mathbf{P}) = \{P_0,\, P_1\}$.
$\mathcal{B}(G)$	The set of all pairs whose first element is a component group label and whose second is a local standard PEPA component in the group specified by that label, e.g. $\mathcal{B}(PR_G(n,\, m)) = \{(\mathbf{P},\, P_0),\, (\mathbf{P},\, P_1),\, (\mathbf{R},\, R_0),\, (\mathbf{R},\, R_1)\}$.
$\mathcal{N}(G)$	The number of all possible local standard PEPA derivative states in each group of G, representing the number of differential equations in the fluid model. Thus $\mathcal{N}(G) = \lvert\mathcal{B}(G)\rvert$. e.g. $\mathcal{N}(PR_G(n,\, m)) = 4$.
$\mathcal{S}(G,\, H)$	The size of the component group with label H. That is, the number of parallel components in the group, e.g. $\mathcal{S}(PR_G(n,\, m),\, \mathbf{P}) = n$.
$\mathcal{S}(G)$	The total size of all component groups in G, e.g. $\mathcal{S}(PR_G(n,\, m)) = n + m$.
$\mathcal{D}_\alpha(G)$	For a given action type, $\alpha \in \mathcal{A}$, the structural depth of a grouped PEPA model, which is the largest number of cooperations involving α, whose immediate effect can be seen by a standard PEPA component enabling an α-action within some component group, e.g. $\mathcal{D}_{task_1}(PR_G(n,\, m)) = 1$. See Appendix A.3 for details.

1.3.2. Deriving ODEs from grouped PEPA models

In this section, we present the fluid translation for PEPA models using the grouped PEPA model framework. We will introduce the following key rate and probability functions based on grouped PEPA model evolution.

$\mathcal{R}_\alpha(G, E, H, P)$	The **component rate function** measures the local rate at which component state P in group H performs an α action in the context of the cooperation within the wider grouped PEPA model G (using counting function E).
$p_\alpha(P, Q)$	The **derivative weighting function** measures the probability that component P evolves to component Q in one α-transition.
$r_\alpha(G, E)$	The **count-oriented apparent rate function** measures the total rate of α being produced by grouped model G (using counting function E).

The quantities which will be subject to the fluid approximation are exposed through an aggregation of a grouped PEPA model's state space. Considering $PR_G(n, m)$ again, we see there are $n \times m$ different ways the initial shared $task_1$ action can be performed because it involves exactly one P_0 and exactly one R_0 component. Each of these transitions occurs at rate:

$$\frac{1}{n}\frac{1}{m}\min(nr_1, mr_2)$$

The aggregation collects states together based on the number of each type of component in each component group. In the case of $PR_G(n, m)$, we might represent the initial aggregate state informally as "$n \times P_0$, $0 \times P_1$, $m \times R_0$ and $0 \times R_1$ components". All of the $n \times m$ transitions above would thus become one transition from the aggregate state "$n \times P_0$, $0 \times P_1$, $m \times R_0$ and $0 \times R_1$ components" to the aggregate state "$(n-1) \times P_0$, $1 \times P_1$, $(m-1) \times R_0$ and $1 \times R_1$ components" at aggregate rate $\min(nr_1, mr_2)$. This aggregation process constructs an *underlying aggregated CTMC*. For example, Figure 2 shows the underlying aggregated CTMC for the 2-processor/2-resource model, $PR_G(2, 2)$.

In general and more formally, it has been shown [11, Theorem 2.12] that the underlying CTMC of a grouped PEPA model can always be aggregated according to the component counts. That is, two states G_1 and $G_2 \in ds(G)$ are aggregated if and only if they have the same number of each type of standard PEPA component in each component group. Then each state of the underlying aggregated CTMC of a grouped PEPA model can be uniquely determined by the model's initial state G and a function $E \in \mathcal{B}(G) \to \mathbb{Z}_+$ which counts the number of standard PEPA components of each type currently active in a given component group. Conversely, note that not all such functions specify valid states in the underlying aggregated CTMC (for example, if it specifies a total number of components in a component group that exceeds the component group's size, or specifies an otherwise unreachable CTMC state).

We may then define the *component rate function* for a grouped PEPA model G, which calculates the aggregate rate at which a standard PEPA component P within a component group H completes an action α, in the aggregate state specified by E. This is needed to describe the rate of evolution of a component group from one derivative state to the next when constructing the fluid model.

Definition 1.1 (Component rate function)**.** *Let G be a grouped PEPA model. For $(H, P) \in \mathcal{B}(G)$, action type $\alpha \in \mathcal{A}$ and $E \in \mathcal{B}(G) \to \mathbb{Z}_+$ specifying the component counts, the component rate is $\mathcal{R}_\alpha(G, E, H, P)$, defined as:*

$$\mathcal{R}_\alpha(M_1 \underset{L}{\bowtie} M_2, E, H, P) := \begin{cases} \frac{\mathcal{R}_\alpha(M_i, E, H, P)}{r_\alpha(M_i, E)} \min(r_\alpha(M_1, E), r_\alpha(M_2, E)) \\ \qquad \textit{if } H \in \mathcal{G}(M_i) \textit{ and } \alpha \in L, \textit{ for } i = 1 \textit{ or } 2 \\ \mathcal{R}_\alpha(M_i, E, H, P) \\ \qquad \textit{if } H \in \mathcal{G}(M_i) \textit{ and } \alpha \notin L, \textit{ for } i = 1 \textit{ or } 2 \end{cases}$$

$$\mathcal{R}_\alpha(Y\{D\}, E, H, P) := \begin{cases} E(H, P)\, r_\alpha(P) & \textit{if } H = Y \textit{ and } P \in \mathcal{B}(G, H) \\ 0 & \textit{otherwise} \end{cases}$$

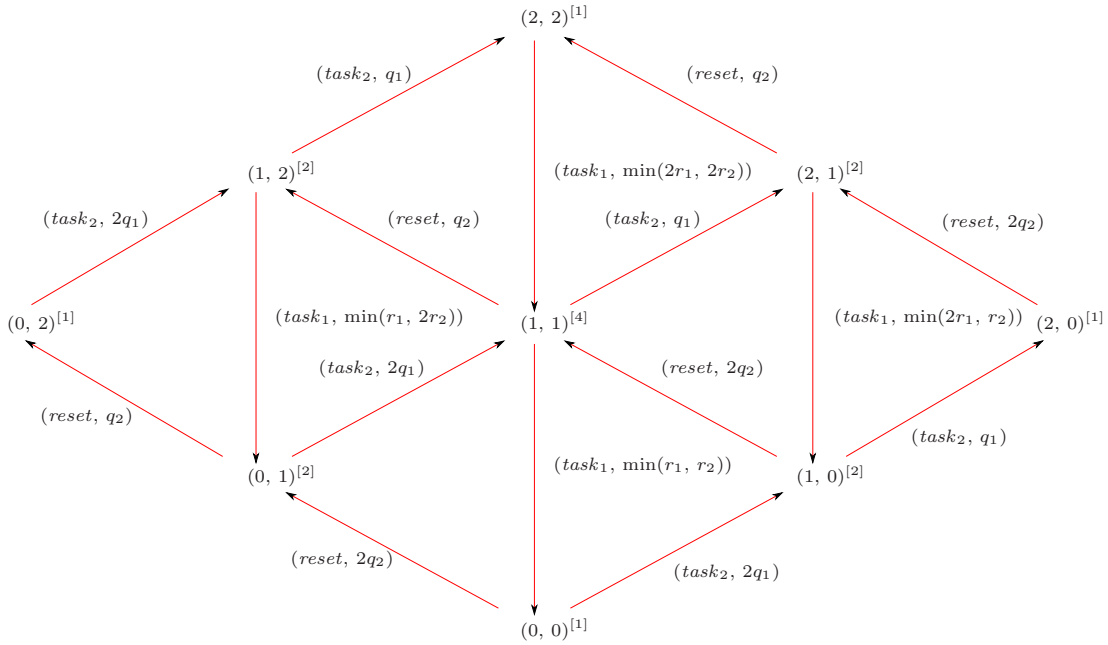


Fig. 2: Underlying aggregated CTMC for simple 2-processor/2-resource model. Each aggregated derivative state is represented by a tuple $(p,\, r)^{[n]}$ where p is the number of P_0 components, and r is the number of R_0 components active in their respective component groups, fully determining the aggregated state of the model. The superscript $[n]$ indicates how many states in the original state space have been merged into this particular aggregated state.

The terms of the form $\frac{\mathcal{R}_\alpha(M_i, E, H, P)}{r_\alpha(M_i, E)} \min(r_\alpha(M_1,\, E),\, r_\alpha(M_2,\, E))$ *are defined as* 0 *when* $r_\alpha(M_i) = 0$.

This definition uses an alternate version of the apparent rate function, defined in terms of component counts, $E \in \mathcal{B}(G) \to \mathbb{Z}_+$. This definition is equivalent to that of Equation (A.1), apart from the explicit specification of component counts by E (hence the prefix *count-oriented*).

Definition 1.2 (Count-oriented apparent rate)**.** *Let G be a grouped PEPA model. Let $\alpha \in \mathcal{A}$ be an action type and $E \in \mathcal{B}(G) \to \mathbb{Z}_+$ specify the component counts. Then the count-oriented apparent rate is $r_\alpha(G,\, E)$, defined as:*

$$r_\alpha(M_1 \underset{L}{\bowtie} M_2,\, E) := \begin{cases} \min(r_\alpha(M_1,\, E),\, r_\alpha(M_2,\, E)) & \text{if } \alpha \in L \\ r_\alpha(M_1,\, E) + r_\alpha(M_2,\, E) & \text{otherwise} \end{cases}$$

$$r_\alpha(Y\{D\},\, E) := \sum_{P \in \mathcal{B}(Y\{D\},\, Y)} E(Y,\, P)\, r_\alpha(P)$$

For example, we have that:

$$\mathcal{R}_{task_1}(PR_G(n,\, m),\, E,\, \mathbf{Processors},\, P_0) = \min(nr_1,\, mr_2) \tag{1.8}$$

assuming $E_0 \in \mathcal{B}(G) \to \mathbb{Z}_+$ represents the initial state of all processors in state P_0 and all resources in state R_0, that is:

$$E_0(\mathbf{Processors},\, P_0) = n \quad E_0(\mathbf{Processors},\, P_1) = 0 \quad E_0(\mathbf{Resources},\, R_0) = m \quad E_0(\mathbf{Resources},\, R_1) = 0$$

For a grouped PEPA model G, let $(H,\, P) \in \mathcal{B}(G)$ and introduce the integer-valued stochastic process, $N_{H,\,P}(t)$, which counts the number of P-components active at a given time $t \geq 0$ within the component

group, H. We intend to define, by means of a system of ODEs, real-valued deterministic functions $v_{H,\,P}(t)$ as approximations to the $N_{H,\,P}(t)$.

The component rate function will be used to define the system of ODEs associated to a grouped PEPA model. In order to support the continuous approximation, we must first however extend the definition of component rate from elements of $\mathcal{B}(G) \to \mathbb{Z}_+$ to elements of $\mathcal{B}(G) \to \mathbb{R}$. This extension is the natural one induced by extending the syntactic definitions (Definitions 1.1 and 1.2) in the obvious manner. Of course, component counts which are not integer-valued have no immediate relationship to the original grouped PEPA model since it makes no sense to have a non-integer number of components. However, this extension is exactly what we need for the fluid approximation, where integer component counts are approximated by real variables.

For some time $t \geq 0$, define $E_t \in \mathcal{B}(G) \to \mathbb{Z}_+$ such that $E_t(H,\,P) = N_{H,\,P}(t)$ for all $(H,\,P) \in \mathcal{B}(G)$. It is clear that E_t represents the aggregated CTMC state at time t. Then it can be shown [11, Theorem 2.15] that $\mathcal{R}_\alpha(G,\,E_t,\,H,\,P)$ is simply the sum of the rates of all outgoing α-transitions from the current aggregated CTMC state to any other, which involves an evolution of a P-component in group H. On the other hand, in order to consider outgoing transitions which involve evolution into a P-component, we need to make one further definition. We define the *derivative weighting function* which calculates the probability that given that a standard PEPA component P does an α-action, when it does so, it transits to another specified standard PEPA component Q.

Definition 1.3 (Derivative weighting function)**.** *Let P and Q be standard PEPA components and let $\alpha \in \mathcal{A}$. Then:*

$$p_\alpha(P,\,Q) := \frac{1}{r_\alpha(P)} \sum_{P \xrightarrow{(\alpha,\,\lambda)} Q} \lambda$$

This is defined to be zero when $r_\alpha(P) = 0$.

Then it is also the case [11, Theorem 2.15] that the sum of the rates of all outgoing α-transitions from the current aggregated CTMC state which involve evolution into a P-component is:

$$\sum_{Q \in \mathcal{B}(G)} p_\alpha(Q,\,P)\,\mathcal{R}_\alpha(G,\,E_t,\,H,\,Q) \tag{1.9}$$

Since the respective terms $p_\alpha(P,\,P)\,\mathcal{R}_\alpha(G,\,E_t,\,H,\,P)$, induced by any self-loops of P to itself, cancel, the rate of all outgoing α-transitions which increase the number of P-components minus the rate of all outgoing α-transitions which decrease the number of P-components is then:

$$\left(\sum_{Q \in \mathcal{B}(G,\,H)} p_\alpha(Q,\,P)\,\mathcal{R}_\alpha(G,\,E_t,\,H,\,Q) \right) - \mathcal{R}_\alpha(G,\,E_t,\,H,\,P) \tag{1.10}$$

Considering the sum of all such terms over all action types then motivates the following definition of the system of ODEs associated to a grouped PEPA model.

Definition 1.4 (ODE system associated with a grouped PEPA model)**.** *Let G be a grouped PEPA model. We define the evolution of the $v_{H,\,P}(t)$ over time for $(H,\,P) \in \mathcal{B}(G)$ by the system of first-order coupled ODEs:*

$$\dot{v}_{H,\,P}(t) = \sum_{\alpha \in \mathcal{A}} \left(\sum_{Q \in \mathcal{B}(G,\,H)} p_\alpha(Q,\,P)\,\mathcal{R}_\alpha(G,\,V(t),\,H,\,Q) \right) - \mathcal{R}_\alpha(G,\,V(t),\,H,\,P)$$

for all $(H,\,P) \in \mathcal{B}(G)$ and where for $t \in \mathbb{R}_+$, $V(t) \in \mathcal{B}(G) \to \mathbb{R}_+$ is given by $V(t) := (\lambda(H,\,P) \to v_{H,\,P}(t))$ for all $(H,\,P) \in \mathcal{B}(G)$. The initial conditions, $V_0 \in \mathcal{B}(G) \to \mathbb{R}_+$ are those naturally defined by the initial state of G.

Note that for non-negative initial conditions, it is immediate from the definition of the ODEs that for any solution, $\dot{v}_{H,P}(t) \geq -v_{H,P}(t)$, and thus, $v_{H,P}(t) \geq 0$ for all $t \in \mathbb{R}_+$. Furthermore since for all $H \in \mathcal{G}(G)$, $\sum_{P \in \mathcal{B}(G,H)} \dot{v}_{H,P}(t) = 0$ and $V_0(H, P) \leq \mathcal{S}(G, H)$, $v_{H,P}(t) \leq \mathcal{S}(G, H)$ for all $t \in \mathbb{R}_+$. That is, any solution to the system of ODEs must at least lie within the natural boundaries imposed by the model they are derived from.

In the general situation of later sections, we will not necessarily wish to carry around so much notation. For a grouped PEPA model G, we can always fix some ordering on the pairs $(H, P) \in \mathcal{B}(G)$, so each $(H, P) \in \mathcal{B}(G)$ corresponds uniquely to some $i \in \{1, \ldots, \mathcal{N}(G)\}$. Accordingly, we may write the system of ODEs of Definition 1.4 simply as $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$, where $\mathbf{v}(t) = (v_1(t), \ldots, v_{\mathcal{N}(G)}(t)) \in \mathbb{R}_+^{\mathcal{N}(G)}$, so that, if i corresponds to (H, P), then $v_i(t) = v_{H,P}(t)$ for all $t \geq 0$. Using the same ordering, write $\mathbf{N}(t)$ as the vector-valued stochastic process with entries, $N_i(t)$ corresponding to each $N_{H,P}(t)$. The following result gives Lipschitz continuity of $\mathbf{f}(\cdot)$, thus guaranteeing the unique existence of a solution to the system of differential equations.

Lemma 1.5. *The system of ODEs, $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$, corresponding to a grouped PEPA model, G, is Lipschitz continuous and a Lipschitz constant is:*

$$\mathcal{K}(G) := 2\mathcal{N}(G) \sum_{\alpha \in \mathcal{A}} (\mathcal{D}_\alpha(G) + 1) \mathcal{Q}_\alpha^{max}(G)$$

Proof. See Appendix C.1. □

Where $\mathcal{Q}_\alpha^{\max}(G)$ and the related function, $\mathcal{Q}^{\max}(G)$ (used in Theorem 3.2), are defined as:

$\mathcal{Q}_\alpha^{\max}(G)$	The **maximal local α-rate function** measures the maximum rate of α actions which can be enabled locally by any standard PEPA component in the grouped model G. So $\mathcal{Q}_\alpha^{\max}(G) := \max_{(H,P) \in \mathcal{B}(G)}\{r_\alpha(P)\}$. e.g. $\mathcal{Q}_{task_1}^{\max}(PR_G(n, m)) = \max\{r_1, r_2\}$.
$\mathcal{Q}^{\max}(G)$	The **maximal local rate function** measures the maximum aggregate rate over all actions which can be enabled locally by any standard PEPA component in the grouped model G. So $\mathcal{Q}^{\max}(G) := \max_{(H,P) \in \mathcal{B}(G)}\{\sum_{\alpha \in \mathcal{A}} r_\alpha(P)\}$. e.g. $\mathcal{Q}^{\max}(PR_G(n, m)) = \max_i\{r_i, q_i\}$.

We will later also require the following straightforward result regarding the system of ODEs. Lemma 1.6 will prove to be fundamental in ensuring that passage-time approximations are comparable for a sequence of structurally-equivalent models. It is will also be used in the convergence proofs of Section 3.

Lemma 1.6. *Let G be a grouped PEPA model. Its corresponding system of ODEs can be written in the form, $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$, as above. For any $\beta \in \mathbb{R}_+$, $\mathbf{f}(\beta\mathbf{x}) = \beta\mathbf{f}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}_+^{\mathcal{N}(G)}$.*

Proof. This follows from the homogeneity of the apparent and component rate functions:

$$r_\alpha(G, \beta E) = \beta r_\alpha(G, E)$$
$$\mathcal{R}_\alpha(G, \beta E, H, P) = \beta \mathcal{R}_\alpha(G, E, H, P)$$

for all $(H, P) \in \mathcal{B}(G)$ and $E \in \mathcal{B}(G) \to \mathbb{R}_+$. Verification takes place using structural induction. □

1.3.3. Fluid analysis example

We now apply Definition 1.4 directly to the simple grouped PEPA model $PR_G(n, m)$, resulting in the following system of ODEs:

$$\begin{aligned}
\dot{v}_{P_0}(t) &= -\min(r_1 v_{P_0}(t), r_2 v_{R_0}(t)) + q_1 v_{P_1}(t) \\
\dot{v}_{P_1}(t) &= \min(r_1 v_{P_0}(t), r_2 v_{R_0}(t)) - q_1 v_{P_1}(t) \\
\dot{v}_{R_0}(t) &= -\min(r_1 v_{P_0}(t), r_2 v_{R_0}(t)) + q_2 v_{R_1}(t) \\
\dot{v}_{R_1}(t) &= \min(r_1 v_{P_0}(t), r_2 v_{R_0}(t)) - q_2 v_{R_1}(t)
\end{aligned} \tag{1.11}$$

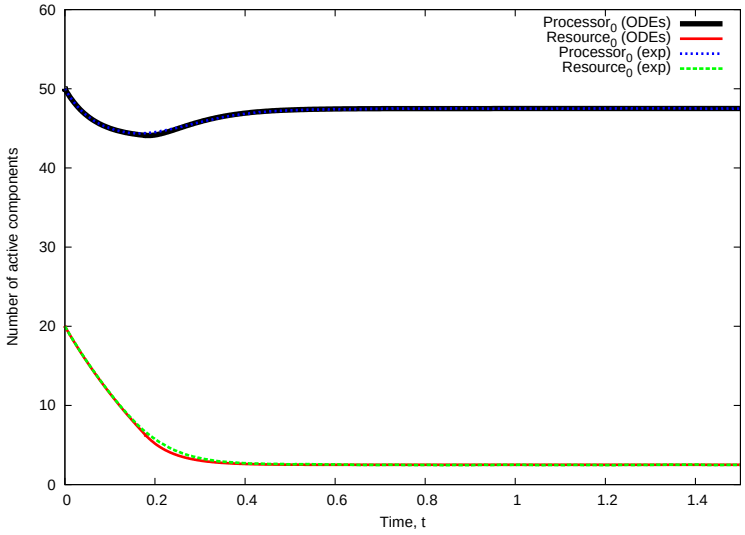


Fig. 3: Comparison of ODE solutions with expectations obtained through stochastic simulation of the underlying CTMC for simple processor/resource model. Rates used are $r_1 = 2.0$, $r_2 = 14.0$, $q_1 = 14.0$ and $q_2 = 2.0$. Initial conditions are 50 P_0 and 20 R_0 components.

where we have abbreviated $v_{\mathbf{Processors},\,P_0}(t)$ as $v_{P_0}(t)$ and $N_{\mathbf{Processors},\,P_0}(t)$ as $N_{P_0}(t)$ and so on. Figure 3 compares the result of integrating these ODEs, with the corresponding expectations, obtained by repeated stochastic simulation of the underlying CTMC. Specifically, we show comparisons between $v_{P_0}(t)$ and $\mathbb{E}[N_{P_0}(t)]$, and $v_{R_0}(t)$ and $\mathbb{E}[N_{R_0}(t)]$.

We observe an impressive correspondence, both in the steady-state and transient phases. Indeed, for much of the time, the ODE solution is indistinguishable from the actual expectation it approximates.

2. Fluid Passage-time Approximations

The purpose of this paper is to show how the fluid approximation techniques introduced in the previous section may be used to compute approximations to passage-time random variables of interest. Bradley *et al.* [2] noted that we can consider certain passage-times as the time-to-extinction of a certain set of components in a modified version of the original model. This exposes the quantities to approximation by fluid-analysis techniques. These ideas were inspired by traditional passage-time analysis techniques [19] where absorbing modifications are made to make passage-time measures more explicit.

For example, in the case of the simple $PR_G(n,\, m)$ model of Equation (1.6), we may be interested in how long it takes for some proportion of the initial n processors to complete their first cycle (consisting of a $task_1$-action followed by a $task_2$-action). As it stands, such a random variable cannot be represented explicitly in the aggregated state space; we wish to represent it as the passage from a given source state to a set of target states. In order to do this, we can modify the $Processor_0$ and $Processor_1$ components and introduce

two new components, $Processor'_0$ and $Processor'_1$, as follows:

$$Processor_0 \stackrel{def}{=} (task_1,\, r_1).Processor_1$$
$$Processor_1 \stackrel{def}{=} (task_2,\, q_1).Processor'_0$$
$$Processor'_0 \stackrel{def}{=} (task_1,\, r_1).Processor'_1$$
$$Processor'_1 \stackrel{def}{=} (task_2,\, q_1).Processor'_0$$

Call the resulting model $PR'_G(n,\, m)$.

$$PR'_G(n,\, m) \stackrel{def}{=} \mathbf{TransientProcessors}\{Processor_0[n]\} \underset{\{task_1\}}{\bowtie} \mathbf{Resources}\{Resource_0[m]\} \tag{2.1}$$

Now we are in a position to express the random variable we are interested in as the time-to-extinction of the specified number of the $Processor_0$ and $Processor_1$ components in the modified model. A central state of the underlying CTMC is shown in Figure 4, where P_i represents the new $Processor_i$ component count in the model $PR'_G(n,\, m)$, similarly P'_i for $Processor'_i$ and R_i for $Resource_i$. Note that it is easy to see how we could develop a similar modification to, for example, allow us to time how long it takes for a processor to complete *any* number of cycles. We will shortly show how the differential equations obtained by applying Definition 1.4 to this modified model (given in Appendix B.1) can be used to compute fluid approximations to such random variables.

In this paper, we will consider two different classes of passage-times, which are particularly amenable to accurate fluid approximation under the right conditions. We will see shortly how the simple framework of the above example actually includes instances of each.

Passage-times of the first type are called *global passage-times*. These passage-times represent the time taken for a significant proportion of a component population to reach some state, or achieve some particular goal.

The second type is called an *individual passage-time*. These will be marginal passage-times for individuals in a large population of identically-distributed components.

2.1. Global Passage-times

We consider again instances of the model introduced above, with even numbers of processors, that is, $PR'_G(2n,\, m)$. Consider the passage-time quantity for half (or n) of the processors completing their initial cycle. As mentioned above, the fluid semantics of Section 1.3.2 can be applied to this model, yielding the system of ODEs of Appendix B.1. In contrast to the case of the unmodified model, $PR_G(2n,\, m)$, however, these ODEs allow us access to the random variable we are interested in. Specifically, it would seem sensible to construct the approximation by considering the deterministic quantity, $v_{P_0}(t) + v_{P_1}(t)$ (or $v_{P'_0}(t) + v_{P'_1}(t)$), which allows us to compute the time, t, at which it reaches the value n. We will shortly present two possible approaches, the first, which yields a deterministic approximation, and the second, which yields approximations to upper and lower bounds on the entire distribution of the passage-time. However, we first define the general class of global passage-times which we will be interested in for fluid analysis.

A *global passage-time* consists of a grouped PEPA model together with an absorbing subset of its aggregated state space, specified by a particular inequality in the components of $\mathbf{N}(t)$, where $N_i(t)$ represents the ith component count at time t, as dictated by some chosen ordering on $(H,\, P) \in \mathcal{B}(G)$ as defined in Section 1.3.2. The passage-time random variable is then the time taken for the model to reach any one of the states in the absorbing subset. The formal definition follows.

Definition 2.1. *Let G be a grouped PEPA model, $\mathbf{c} \in \{0,\, 1\}^{\mathcal{N}(G)}$ be a vector which selects components in $\mathbf{N}(t)$ and $C \in \mathbb{Z}_+$, which represents the target component count. Define the global passage-time random variable, $\sigma := \inf\{t \in \mathbb{R}_+ : \mathbf{c} \cdot \mathbf{N}(t) \leq C\}$, where whenever $t > \sigma$, $\mathbf{c} \cdot \mathbf{N}(t) \leq C$, that is, the target states are absorbing.*

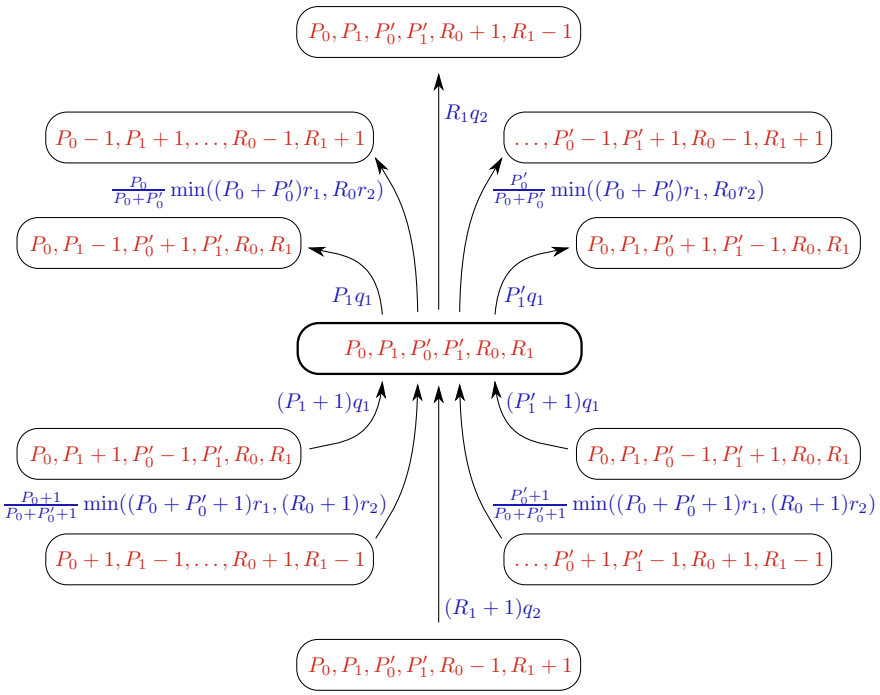


Fig. 4: A central state of the underlying aggregated CTMC of the model $PR'_G(n, m)$

This definition can be used to describe the example passage-time for half of the processors to complete a cycle by letting $G = PR'_G(2n, m)$, $\mathbf{c} = (1, 1, 0, 0, 0, 0)$ and $C = n$, assuming an ordering of the vector $\mathbf{v}(t)$, such that $v_{P_0}(t) \equiv v_1(t)$ and $v_{P_1}(t) \equiv v_2(t)$, and similarly for $\mathbf{N}(t)$.

2.1.1. Deterministic approximations

The most straightforward approach to approximating the passage-time mentioned above would be to compute the time, t, at which the quantity, $v_{P_0}(t) + v_{P_1}(t)$, reaches n.

Figure 5 shows probability density functions computed using traditional methods for this passage-time random variable. In each case, we increase the number of processors and there are always half as many resources as there are processors. Maintaining this ratio ensures that the deterministic approximation for each of these passage-times is actually the same (Lemma 1.6), represented by γ in the figure. We see that the probability density functions do appear to be converging to the point mass at γ as the component populations increase.

For a general global passage-time, specified for some grouped PEPA model, G, by $\mathbf{c} \in \{0, 1\}^{\mathcal{N}(G)}$ and $C \in \mathbb{Z}_+$, as in Definition 2.1, the deterministic approximation, γ, is defined simply as:

$$\gamma := \inf\{t \in \mathbb{R}_+ \ : \ \mathbf{c} \cdot \mathbf{v}(t) \leq C\} \tag{2.2}$$

In Section 3, we will show that the limiting result depicted in Figure 5 holds in general.

2.1.2. Distribution approximations

As can be seen from Figure 5, models with smaller component populations (say, $n \leq 128$ in this example) generate passage-time densities with significant variance. In this case, a deterministic passage-time result does not adequately capture the full picture. In this section, we show the CDF of the passage-time distribution can be bounded for smaller component population sizes.

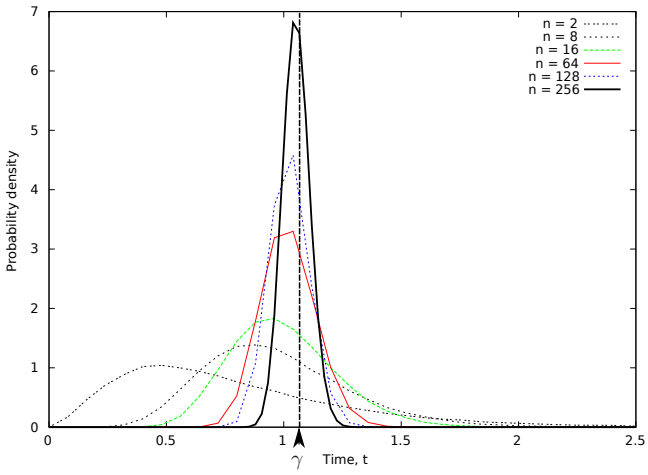


Fig. 5: Probability density functions of passage-time for half of the processors to complete their first two tasks in the model $PR'_G(2n, n)$ compared with the ODE approximation, denoted γ. Rates: $r_1 = 2.0$, $r_2 = 2.0$, $q_1 = 8.0$ and $q_2 = 3.0$.

We approximate global passage-time distributions by employing the well-known Markov inequality, which says that for a non-negative random variable X, and $a > 0$:

$$\mathbb{P}\{X \geq a\} \leq \frac{\mathbb{E}[X]}{a} \tag{2.3}$$

In order to exploit this, we use the observation from [11] that $\mathbb{E}[\mathbf{N}(t)]$ is often well approximated by $\mathbf{v}(t)$.

Consider again the model $PR'_G(2n, m)$ and the passage-time for n of the processors to complete their first cycle. Denote this random variable by σ, then applying Markov's inequality, we may obtain the following bounds on its cumulative distribution function:

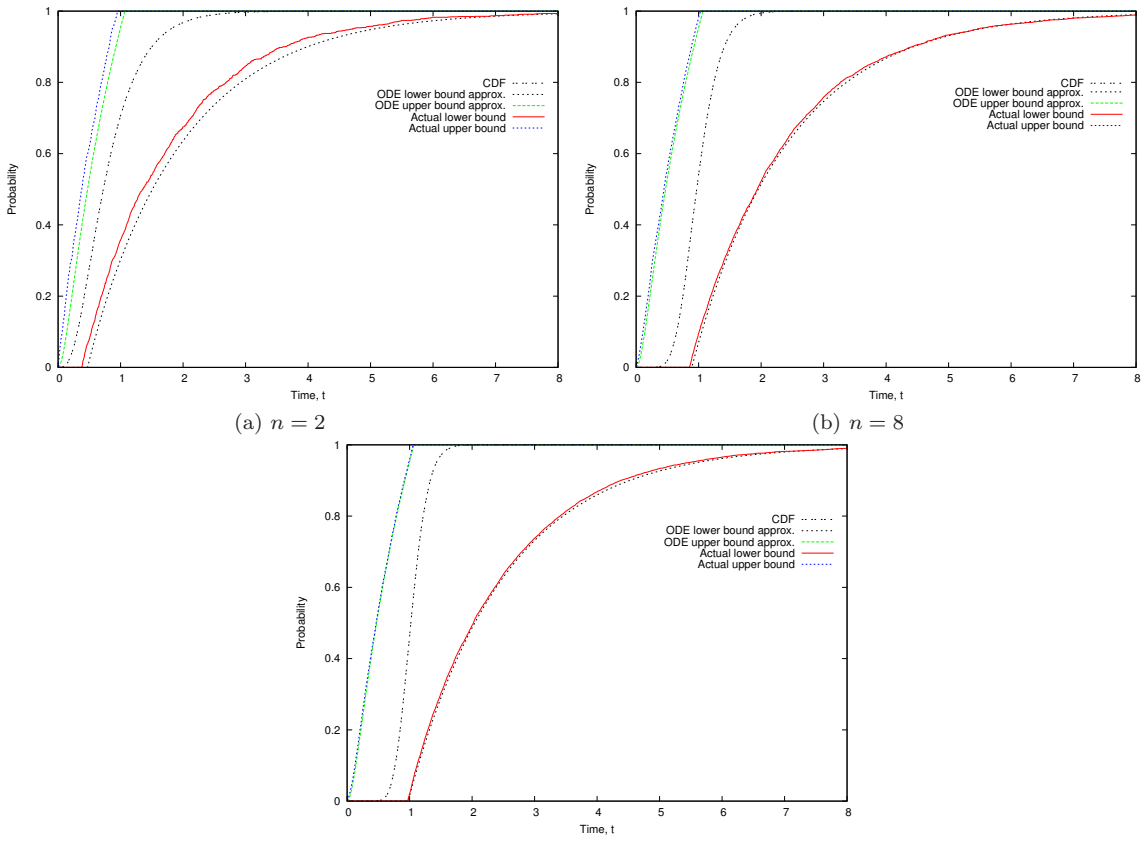
$$\mathbb{P}\{\sigma \leq t\} = \mathbb{P}\{N_{P'_0}(t) + N_{P'_1}(t) \geq n\} \leq \frac{\mathbb{E}[N_{P'_0}(t)] + \mathbb{E}[N_{P'_1}(t)]}{n} \tag{2.4}$$

and:

$$\mathbb{P}\{\sigma \leq t\} = 1 - \mathbb{P}\{N_{P_0}(t) + N_{P_1}(t) \geq n+1\} \geq 1 - \frac{\mathbb{E}[N_{P_0}(t)] + \mathbb{E}[N_{P_1}(t)]}{n+1} \tag{2.5}$$

Applying the approximation $\mathbb{E}[\mathbf{N}(t)] \approx \mathbf{v}(t)$ allows us to estimate these bounds using the solutions to the corresponding system of differential equations. Figure 6 shows cumulative distribution functions for the passage-time random variable for three different processor/resource configurations, with the same ratio of processors to resources. We also plot for each combination, the actual bounds of Equations (2.4) and (2.5) with their differential equation approximations.

The key point to note when comparing Figures 5 and 6 is that, in Figure 6, the actual bound is converging to the approximate bound much faster than the convergence of the probability density function to the point mass in Figure 5. This is to be expected because the previous section required the entire distribution to concentrate around the point predicted by the differential equation solution, whereas in this section, we require only a convergence of expectations. Therefore, it would seem that the distribution bound approximations of this section may be more useful for smaller component populations than using the deterministic approximation of the previous section. This is an improvement paid for by the fact that only bounds on the cumulative distribution function are obtained. However, Figure 6 also indicates that as the population gets larger, the bounds can become quite loose, so at some point, it is certainly likely to be advantageous



(a) $n = 2$ (b) $n = 8$

(c) $n = 16$

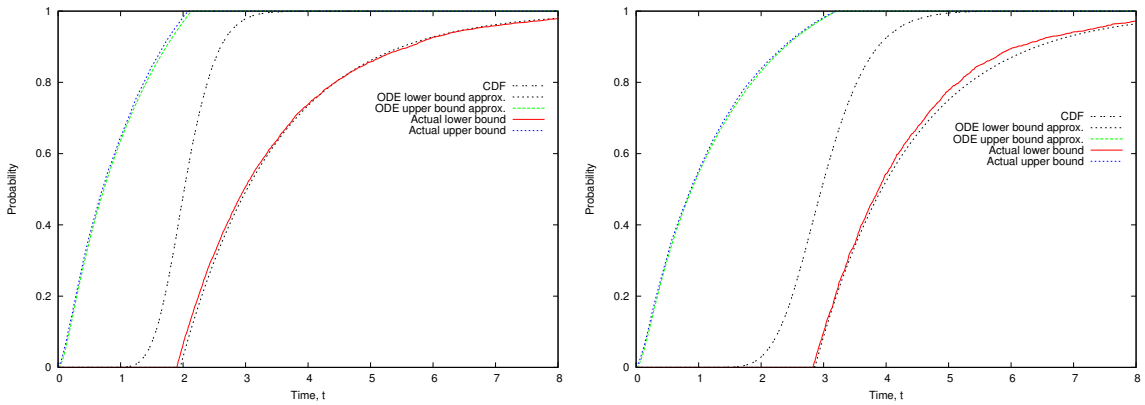
Fig. 6: Cumulative distribution functions of passage-time for half of the processors to complete their first two tasks in the model $PR'_G(2n, n)$ compared with Markov inequality bounds and their ODE approximations. Rates: $r_1 = 2.0$, $r_2 = 2.0$, $q_1 = 8.0$ and $q_2 = 3.0$.

to switch to the deterministic approximation. Indeed, we would not expect that the bounds would become tighter in the limit of large populations, only that the differential equation approximation to the bounds would improve.

Figure 7 shows a similar set of results, but this time we consider only one model ($PR'_G(32, 16)$, that is, $n = 16$ in Figure 6) and vary the passage-time measure we are calculating. We compute the passage-time cumulative distribution function for three quarters (24), seven eighths (28) and finally, all[1] 32 of the processors in this model to complete a cycle. We also compute the exact Markov bounds and their differential equation approximations. The interesting feature of note here is that the relative tightness of the lower bound appears to be increasing in the higher and arguably more useful quantiles as the proportion being timed increases, whereas the upper bound becomes looser everywhere. For $1 \leq a \leq 32$, let σ_a be the passage-time random variable for a of the 32 processors to complete a cycle, then the Markov lower bound inequality can be derived, as above, for each a:

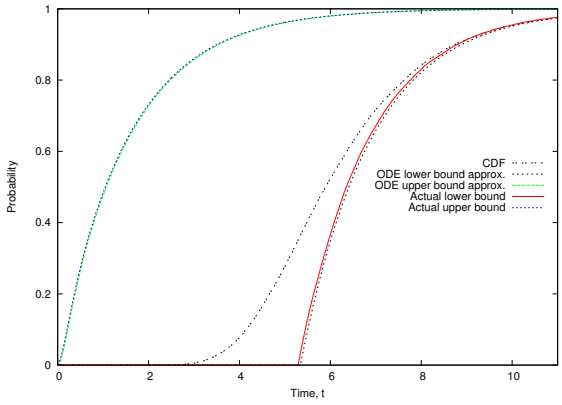
$$\mathbb{P}\{\sigma_a \leq t\} \geq 1 - \frac{1}{a+1}(\mathbb{E}[N_{P_0}(t)] + \mathbb{E}[N_{P_1}(t)]) \tag{2.6}$$

[1]Note that since no component of the differential equation solution ever actually reaches zero, the deterministic approximation of the last section cannot be applied to measure such times to complete extinction of a class of component types.



(a) Three quarters (24)

(b) Seven eighths (28)



(c) All (32)

Fig. 7: Cumulative distribution functions of passage-time for increasing proportions of the processors to complete their first two tasks in the model $PR'_G(32,\, 16)$ compared with Markov inequality bounds and their ODE approximations. Rates: $r_1 = 2.0$, $r_2 = 2.0$, $q_1 = 8.0$ and $q_2 = 3.0$.

This can be rewritten as $\frac{1}{a+1}(\mathbb{E}[N_{P_0}(t)] + \mathbb{E}[N_{P_1}(t)]) \geq \mathbb{P}\{\sigma_a > t\} + 0 \times \mathbb{P}\{\sigma_a \leq t\}$. In higher quantiles, the difference in the two sides of this inequality due to the $0 \times \mathbb{P}\{\sigma_a \leq t\}$ term dominates and this could be expected to decrease with a. Similar considerations can be made for the behaviour of the upper bound as a is varied.

We now consider how to bound global passage-time random variables in general. Let σ be a global passage-time random variable, specified for some grouped PEPA model, G, by $\mathbf{c} \in \{0,\, 1\}^{\mathcal{N}(G)}$ and $C \in \mathbb{Z}_+$, as in Definition 2.1. Write, for $i \in \{1,\, \ldots,\, \mathcal{N}(G)\}$, c_i, for the elements of $\mathbf{c}$. Recall that each i corresponds to an element $(H,\, P) \in \mathcal{B}(G)$, and we write c_i or $c_{H,\,P}$ interchangeably, as in the case of $N_i(t)$ and $N_{H,\,P}(t)$. We may then compute, writing $\phi(G) := \{H \,:\, H \in \mathcal{G}(G), P \in \mathcal{B}(G,\, H), c_{H,\,P} = 1\}$ for the component groups involved in the specification of the global passage time:

$$\begin{aligned}\mathbb{P}\{\sigma \leq t\} = \mathbb{P}\{\mathbf{c} \cdot \mathbf{N}(t) \leq C\} &= \mathbb{P}\left\{\sum_{H \in \phi(G)} \sum_{P \in \mathcal{B}(G,\, H)} c_{H,\,P} \times N_{H,\,P}(t) \leq C\right\} \\ &= \mathbb{P}\left\{\sum_{H \in \phi(G)} \left(\mathcal{S}(G,\, H) - \sum_{P \in \mathcal{B}(G,\, H)} c_{H,\,P} \times N_{H,\,P}(t)\right) \geq \sum_{H \in \phi(G)} \mathcal{S}(G,\, H) - C\right\}\end{aligned} \tag{2.7}$$

Since for any passage-time not identically zero, the right-hand side of the above is strictly positive, we may apply Markov's inequality to obtain:

$$\mathbb{P}\{\sigma \leq t\} \leq \frac{\sum_{H \in \phi(G)} (\mathcal{S}(G,\, H) - \sum_{P \in \mathcal{B}(G,\, H)} c_{H,\, P} \times \mathbb{E}[N_{H,\, P}(t)])}{\sum_{H \in \phi(G)} \mathcal{S}(G,\, H) - C} \tag{2.8}$$

Working in the other direction, we have:

$$\mathbb{P}\{\sigma \leq t\} = 1 - \mathbb{P}\{\mathbf{c} \cdot \mathbf{N}(t) > C\} = 1 - \mathbb{P}\{\mathbf{c} \cdot \mathbf{N}(t) \geq C + 1\} \tag{2.9}$$

Applying Markov's inequality directly, we obtain:

$$\mathbb{P}\{\sigma \leq t\} \geq 1 - \frac{1}{C+1} \sum_{(P,\, H) \in \mathcal{B}(G)} c_{H,\, P} \times \mathbb{E}[N_{H,\, P}(t)] \tag{2.10}$$

The approximation $\mathbb{E}[\mathbf{N}(t)] \approx \mathbf{v}(t)$ can then be applied directly in either case to provide the bound estimates. In Section 3, we will show that the limiting result depicted in Figure 6 holds in general, that is, that the differential equation estimates converge to the actual bounds in the limit of large populations.

2.2. Individual Passage-times

We now define a second type of passage-time measure called individual passage-times, which are amenable to fluid approximation. These are marginal passage-times for individuals in a large population of identically-distributed components.

Consider again the model $PR'_G(n,\, m)$ above and assume we are now interested in how long it takes for any one of the initial n processors to complete one cycle. Since all members of the **TransientProcessors** component group are identically distributed, it makes no difference which specific individual we consider. We will see that this is a necessary requirement for what follows.

Let $I_j(t)$ for $1 \leq j \leq n$ be the stochastic processes which tracks the state of the jth individual in the **TransientProcessors** component group. Note that these processes are identically distributed. Then, we wish to evaluate at time $t \in \mathbb{R}_+$ and for any $1 \leq k \leq n$, the cumulative distribution function:

$$\mathbb{P}\{I_k(t) \neq P_0 \text{ and } I_k(t) \neq P_1\} \tag{2.11}$$

That is, the probability that by time t, a specific individual processor has completed one cycle. Now note that for any $1 \leq k \leq n$:

$$\mathbb{E}[N_{P_0}(t)] = \sum_{j=1}^{n} \mathbb{E}[\mathbf{1}_{\{I_j(t) = P_0\}}] = n\mathbb{P}\{I_k(t) = P_0\} \tag{2.12}$$

So, we could compute the quantity of interest as:

$$\mathbb{P}\{I_k(t) \neq P_0 \text{ and } I_k(t) \neq P_1\} = 1 - \frac{\mathbb{E}[N_{P_0}(t)] + \mathbb{E}[N_{P_1}(t)]}{n} \tag{2.13}$$

If we now approximate $\mathbb{E}[N_{P_0}(t)]$ and $\mathbb{E}[N_{P_1}(t)]$ by $v_{P_0}(t)$ and $v_{P_1}(t)$, respectively, this provides a route to the fluid approximation of the entire cumulative distribution function of the individual passage-times.

Figure 8 shows cumulative distribution functions computed using traditional methods for the passage-time random variable discussed above. In each case, we increase the number of processors and there are always a quarter as many resources as there are processors. This ensures, similarly to the previous section, that the differential equation approximation to the distribution for each of these passage-times is actually the same (Lemma 1.6). We see that the cumulative distribution functions do appear to be converging to the fluid approximation as the component populations increase.

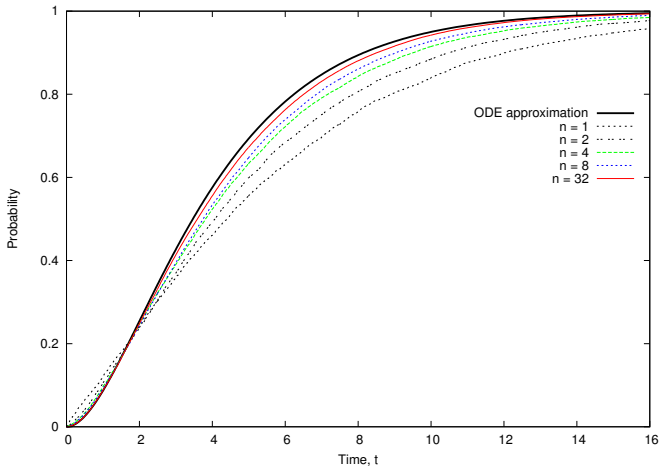


Fig. 8: Cumulative distribution functions (CDFs) of passage-time for one processor to complete its first two tasks in the model $PR'_G(4n,\, n)$ compared with the ODE approximation to the CDF. Rates: $r_1 = 0.4$, $r_2 = 6.0$, $q_1 = 0.6$ and $q_2 = 1.3$.

The general class of individual passage-times we will be interested in for fluid analysis is now given exactly, as we did for global passage-times. An individual passage-time consists of a grouped PEPA model, which has a component group, all of whose components start from the same initial state, so that they are identically-distributed. Furthermore, some absorbing set of local target states for this component must be specified to determine the destination of the passage being timed. The formal definition follows.

Definition 2.2. *Assume we are given:*

- *A grouped PEPA model, G*
- *A component group $H \in \mathcal{G}(G)$, such that all components within H start in the same initial state, that is, for some standard PEPA component $P \in \mathcal{B}(G,\, H)$, $N_{H,\,P}(0) = \mathcal{S}(G,\, H)$*
- *A set of local target states of P, $T \subseteq ds(P)$, which is absorbing in the sense that $ds(T) \subseteq T$*

Let $I_P(t) \in ds(P)$ be the stochastic process tracking the state of any one of these initial P components in component group H. Then an individual passage-time is defined by the random variable, $\theta := \inf\{t \in \mathbb{R}_+ : I_P(t) \in T\}$.

This definition can be used to describe the example passage-time discussed above with $G = PR'_G(n,\, m)$, $H =$ **TransientProcessors** and $T = \{P'_0,\, P'_1\}$. We believe it should also be possible to consider joint passage times for pairs of components (and other combinations) by considering the individual passage times that can be obtained from models such as:

$$\textbf{TransientProcessors}\{(Processor_0 \,||\, Processor_0)[n/2]\} \underset{\{task_1\}}{\bowtie} \textbf{Resources}\{Resource_0[m]\} \tag{2.14}$$

In the case of a general individual passage-time, we have:

$$\mathbb{P}\{\theta \leq t\} = \mathbb{P}\{I_P(t) \in T\} = \frac{1}{\mathcal{S}(G,\, H)} \sum_{Q \in T} \mathbb{E}[N_{H,\,Q}(t)] \tag{2.15}$$

The approximation $\mathbb{E}[\mathbf{N}(t)] \approx \mathbf{v}(t)$ can then be applied directly to provide the approximation to the cumulative distribution function.

3. Limiting Convergence of Approximations

In this section, we present results which give convergence of the fluid approximations for both global and individual passage-times in sequences of grouped PEPA models for increasing total component population. The proof of Theorem 3.2 below is based on the decomposition of the underlying aggregated CTMC into the sum of a martingale and a predictable process. The predictable process is guaranteed unique by the Doob–Meyer decomposition theorem and is usually called the *compensator* [20, Chap. 25]. The methodology of using this decomposition to prove this kind of limit result was originally developed by Darling *et al.* [21, 22]. For the technical details of the following discussion, it is important to note that we are considering the construction of the aggregated CTMC, $\mathbf{N}(t)$, of a grouped PEPA model, to be such that the traces are all right continuous with left limits (*càdlàg*).

The fluid limit results arise through the rescaling of a grouped PEPA model's underlying aggregated CTMC by the total component population size. Specifically, for a grouped PEPA model, G, we will be interested in the rescaled process:

$$\bar{\mathbf{N}}(t) := \frac{1}{\mathcal{S}(G)}\mathbf{N}(t) \tag{3.1}$$

and its fluid approximation:

$$\bar{\mathbf{v}}(t) := \frac{1}{\mathcal{S}(G)}\mathbf{v}(t) \tag{3.2}$$

So $\bar{N}_{H,\,P}(t)$ is the proportion of the total component population, which are in group H and state P at time t. Note that Lemma 1.6 tells us that the rescaled quantities, $\bar{\mathbf{v}}(t)$, satisfy the same system of differential equations as the unscaled ones, that is, $\dot{\bar{\mathbf{v}}}(t) = \mathbf{f}(\bar{\mathbf{v}}(t))$, of course with scaled initial conditions. From this point onwards, we will work primarily with the rescaled process and its rescaled fluid approximation.

Define now the stochastic process:

$$\bar{\mathbf{M}}(t) := \bar{\mathbf{N}}(t) - \left(\bar{\mathbf{N}}(0) + \int_0^t \mathbf{f}(\bar{\mathbf{N}}(s))\,ds\right) \tag{3.3}$$

where $\mathbf{f}(\cdot)$ is the rate at which components counts are incremented minus that at which they are decremented, also defined in the previous section. We now show that this is indeed the decomposition guaranteed by Doob–Meyer by showing that $\bar{\mathbf{M}}(t)$ is a martingale, so that the predictable process $\bar{\mathbf{A}}(t) := \bar{\mathbf{N}}(0) + \int_0^t \mathbf{f}(\bar{\mathbf{N}}(s))\,ds$ is the compensator of $\bar{\mathbf{N}}(t)$.

Theorem 3.1. *The process $\bar{\mathbf{M}}(t)$ is a vector martingale.*

Proof. See Appendix C.2. □

We now use this decomposition to obtain a probabilistic estimate on the magnitude of the difference between $\bar{\mathbf{N}}(t)$ and $\bar{\mathbf{v}}(t)$. This is the key result which will allow us to show the convergence of the fluid passage-time approximations as suggested empirically in the last section.

Theorem 3.2. *Let G be a grouped PEPA model, with associated rescaled aggregated stochastic process, $\bar{\mathbf{N}}(t)$ and fluid approximation, $\bar{\mathbf{v}}(t)$. Fix some $T > 0$ and $\epsilon > 0$. Then:*

$$\mathbb{P}\left\{\sup_{t\in[0,\,T]} \|\bar{\mathbf{N}}(t) - \bar{\mathbf{v}}(t)\| > \epsilon\right\} \leq \frac{\gamma(\mathcal{S}(G)qT,\, \mathcal{S}(G)\mathcal{Q}^{max}(G)T)}{(\mathcal{S}(G)qT-1)!} + \frac{4qT\mathcal{N}(G)}{\mathcal{S}(G)\epsilon^2}\exp(2\mathcal{K}(G)T)$$

where q can be any positive real number strictly greater than $\mathcal{Q}^{max}(G)$ and chosen such that qT is an integer, for example, $q = \frac{\lceil \mathcal{Q}^{max}(G)T\rceil+1}{T}$.

Furthermore, if all of q, T, $\mathcal{Q}^{max}(G)$, $\mathcal{N}(G)$, ϵ and $\mathcal{K}(G)$ remain fixed, and $\mathcal{S}(G) \to \infty$, this bound tends to zero, and thus, so does $\mathbb{P}\{\sup_{t\in[0,\,T]} \|\bar{\mathbf{N}}(t) - \bar{\mathbf{v}}(t)\| > \epsilon\}$.

Proof. See Appendix C.3. □

The convergence of passage-times will be as the total component population size increases in a sequence of *structurally equivalent* grouped PEPA models. We define this notion in the next section before proceeding to give the passage-time convergence results.

3.1. Structurally Equivalent Grouped PEPA Models

When two grouped PEPA models are structurally the same, differing only in that they may have different component population sizes, but in the same ratios, we say that they are *structurally equivalent*.

Definition 3.3 (Structural equivalence)**.** *Let G_1 and G_2 be two grouped PEPA models. Then we say they are structurally equivalent if firstly, they have the same model structure, $\mathcal{B}(G_1) = \mathcal{B}(G_2) =: \mathcal{B}$ and $\mathcal{S}(G_1, G_2) = \mathbf{true}$, where $\mathcal{S}(\cdot, \cdot)$ is defined as:*

$$\mathcal{S}(M_1 \underset{L}{\bowtie} M_2, N_1 \underset{L}{\bowtie} N_2) := \mathcal{S}(M_1, N_1) \wedge \mathcal{S}(M_2, N_2)$$
$$\mathcal{S}(Y\{D_1\}, Y\{D_2\}) := \mathbf{true}$$

and **false** *in all other cases. Secondly, they must have the same initial component population ratios, that is, for all $(H, P) \in \mathcal{B}$, $N^{G_1}_{H,P}(0)/\mathcal{S}(G_1) = N^{G_2}_{H,P}(0)/\mathcal{S}(G_2)$. $N^{G_1}_{H,P}(t)$ resp. $N^{G_2}_{H,P}(t)$ is the stochastic process counting the number of P components in group H in the model G_1 resp. G_2.*

When we say a *set* or *sequence* of grouped PEPA models is structurally equivalent, we mean that each pair in it is. We have already considered a few such sequences, one example is, $\{PR'(2n, n)\}_{n=1}^{\infty}$.

Structural equivalence preserves a number of properties, which is something we will call on in the next section. Before proceeding, we summarise these properties briefly; the following statements hold for any two structurally equivalent grouped PEPA models, say, G_1 and G_2.

- They have the same set of component groups ($\mathcal{G}(G_1) = \mathcal{G}(G_2)$) and the same set of component group and derivative state pairs ($\mathcal{B}(G_1) = \mathcal{B}(G_2)$);
- They have the same maximal local rate and, for any action type $\alpha \in \mathcal{A}$, the same maximal local α rate ($\mathcal{Q}^{\max}(G_1) = \mathcal{Q}^{\max}(G_2)$ and $\mathcal{Q}^{\max}_{\alpha}(G_1) = \mathcal{Q}^{\max}_{\alpha}(G_2)$);
- For any action type, $\alpha \in \mathcal{A}$, they have the same structural depth ($\mathcal{D}_{\alpha}(G_1) = \mathcal{D}_{\alpha}(G_2)$);
- They have the same system of differential equations, (Definition 1.4), just with different initial conditions (in the same ratios);
- The Lipschitz constant given by Lemma 1.5 is the same for both G_1 and G_2 and is thus a Lipschitz constant for both of their systems of differential equations.

3.2. Convergence of Passage-time Approximations

We now present the convergence results for global and individual passage-times to their differential equation approximations.

3.2.1. Global passage-times

For a given sequence of structurally equivalent grouped PEPA models, $\{G_i\}_{i=1}^{\infty}$, write $\mathcal{N} := \mathcal{N}(G_i)$ for any i, which is well-defined by structural equivalence. Let $\mathbf{N}^{G_i}(t)$ be the underlying aggregated CTMC for G_i. Fix some $\mathbf{c} \in \{0,\, 1\}^{\mathcal{N}}$ and $C \in \mathbb{Q}_+ \cap [0,\, 1]$, such that $C \times \mathcal{S}(G_i) \in \mathbb{Z}_+$ for all i. Then define:

$$\sigma_i := \inf\{t \in \mathbb{R}_+ \,:\, \mathbf{c} \cdot \mathbf{N}^{G_i}(t) \leq \mathcal{S}(G_i) \times C\} \tag{3.4}$$

If, for each i, whenever $t > \sigma_i$, $\mathbf{c} \cdot \mathbf{N}^{G_i}(t) \leq \mathcal{S}(G_i) \times C$, then $\{\sigma_i\}_{i=1}^{\infty}$ is a sequence of global passage-times in accordance with Definition 2.1. For example, if we consider again the structurally equivalent sequence of grouped PEPA models, $\{PR'(2n,\, n)\}_{n=1}^{\infty}$ and let $\mathbf{c} = (1,\, 1,\, 0,\, 0,\, 0,\, 0)$ and $C = 1/3$, we obtain a sequence of global passage-times. Indeed, some elements of this sequence are shown in Figure 5 along with their differential equation approximation, γ.

We will now proceed to show in general that the σ_i converge in probability (and thus in distribution) to the deterministic quantity, γ, as suggested empirically in Section 2.1.1.

By structural equivalence, the differential equation approximation for each i is $\mathbf{v}^{G_i}(t)$, where $\dot{\mathbf{v}}^{G_i}(t) = \mathbf{f}(\mathbf{v}^{G_i}(t))$, for the same function, $\mathbf{f}(\cdot)$, independent of i. Furthermore, by homogeneity (Lemma 1.6), the rescaled quantities, $\bar{\mathbf{v}}^{G_i}(t) := \mathbf{v}^{G_i}(t)/\mathcal{S}(G_i)$ also satisfy the same differential equation. But, by structural equivalence, the initial conditions for this differential equation are the same, independent of i. So, by uniqueness of the differential equation solution (Lemma 1.5), the rescaled approximation is independent of i, and we write just $\bar{\mathbf{v}}(t)$. Thus the deterministic approximation to σ_i, say γ_i, defined in Section 2.1.1, can be stated independently of i:

$$\gamma_i := \inf\{t \in \mathbb{R}_+ \,:\, \mathbf{c} \cdot \mathbf{v}^{G_i}(t) \leq \mathcal{S}(G_i) \times C\} = \inf\{t \in \mathbb{R}_+ \,:\, \mathbf{c} \cdot \bar{\mathbf{v}}(t) \leq C\} =: \gamma \tag{3.5}$$

The following theorem then gives the desired convergence result.

Theorem 3.4. *Fix a sequence of structurally equivalent grouped PEPA models, $\{G_i\}_{i=1}^{\infty}$, with all of the above notation. Fix also some $\mathbf{c} \in \{0,\, 1\}^{\mathcal{N}}$ and $C \in \mathbb{Q}_+ \cap [0,\, 1]$, such that $C \times \mathcal{S}(G_i) \in \mathbb{Z}_+$ for all i, so that the sequence $\{\sigma_i\}_{i=1}^{\infty}$ is a sequence of global passage-times and write γ for their deterministic approximation, all as above.*

Assume that $\gamma < \infty$ and further that for all $t > \gamma$, $\mathbf{c} \cdot \bar{\mathbf{v}}(t) < C$. Then, if $\mathcal{S}(G_i) \to \infty$ as $i \to \infty$, we have for any $\epsilon > 0$, $\mathbb{P}\{|\sigma_i - \gamma| > \epsilon\} \to 0$ as $i \to \infty$.

Proof. See Appendix C.4. □

One example of this theorem is then the convergence illustrated in Figure 5.

Now we turn to the approximation illustrated in Section 2.1.2, where we defined estimates for bounds on the cumulative distribution functions of global passage-times, obtained using Markov's inequality. The following theorem shows that in the same limiting scenario as above, these estimates converge to the actual bounds. As in Section 2.1.2, we use the notation $\phi(G) := \{H \,:\, H \in \mathcal{G}(G), P \in \mathcal{B}(G,\, H), c_{H,\, P} = 1\}$ for the component groups involved in the specification of some global passage time specified partially by $\mathbf{c}$.

Theorem 3.5. *Fix a sequence of structurally equivalent grouped PEPA models, $\{G_i\}_{i=1}^{\infty}$, with all of the above notation. Fix also some $\mathbf{c} \in \{0,\, 1\}^{\mathcal{N}}$ and $C \in \mathbb{Q}_+ \cap (0,\, 1]$, such that $C \times \mathcal{S}(G_i) \in \mathbb{Z}_+$ for all i, so that the sequence $\{\sigma_i\}_{i=1}^{\infty}$ is a sequence of global passage-times.*

Fix $T > 0$. Assume that $\mathcal{S}(G_i) \to \infty$ as $i \to \infty$. Then, for all $t \in [0,\, T]$:

$$\left| \frac{\sum_{H \in \phi(G_i)} (\mathcal{S}(G_i,\, H) - \sum_{P \in \mathcal{B}(G_i,\, H)} c_{H,\, P} \times \mathbb{E}[N^{G_i}_{H,\, Q}(t)])}{\sum_{H \in \phi(G_i)} \mathcal{S}(G_i,\, H) - C \times \mathcal{S}(G_i)} - \frac{\sum_{H \in \phi(G_i)} (\mathcal{S}(G_i,\, H) - \sum_{P \in \mathcal{B}(G_i,\, H)} v^{G_i}_{H,\, Q}(t))}{\sum_{H \in \phi(G_i)} \mathcal{S}(G_i,\, H) - C \times \mathcal{S}(G_i)} \right| \longrightarrow 0$$

and

$$\left| \frac{1}{C \times \mathcal{S}(G_i) + 1} \sum_{(H,\, P) \in \mathcal{B}(G_i)} c_{H,\, P} \times \mathbb{E}[N^{G_i}_{H,\, Q}(t)] - \frac{1}{C \times \mathcal{S}(G_i) + 1} \sum_{(H,\, P) \in \mathcal{B}(G_i)} c_{H,\, P} \times v^{G_i}_{H,\, Q}(t) \right| \longrightarrow 0$$

as $i \to \infty$. That is, the approximate bounds of Section 2.1.2 converge to the actual bounds as $i \to \infty$.

Proof. See Appendix C.5. □

An illustration of this theorem is Figure 6. Notice that we have only proved the above theorem for $C \neq 0$, that is, it does not necessarily hold for sequences of global passage-times which measure the time to total component extinction of a class of components. Even so, Figure 7c shows Markov bounds for such a case with very impressive differential equation approximations.

As discussed in Section 2.1.2, the reason why both Theorems 3.4 and 3.5 are useful is that we would expect the convergence to happen faster in the latter case than in the former case. This is because the latter demands only a convergence of expectations whereas the former demands concentration of measure around a deterministic point. This is also justified empirically by the examples in both Sections 2 and 4.

3.2.2. Individual passage-times

We turn now to proving a convergence result for individual passage-times. Similarly to the previous section, for a given sequence of structurally equivalent grouped PEPA models, $\{G_i\}_{i=1}^{\infty}$, write $\mathcal{G} := \mathcal{G}(G_i)$ for any i, which is well-defined by structural equivalence. Now fix some component group $H \in \mathcal{G}$ such that for any i, the initial local state of all components in group H is some standard PEPA component, P. Furthermore, let $T \subseteq ds(P)$ be some absorbing set of local target states, that is, $ds(T) \subseteq T$. Let $I^i_P(t) \in ds(P)$ be the stochastic process tracking the state of any one of these initial P components (recall they are identically-distributed) in component group H in the model G_i. Also, let $\mathbf{N}^{G_i}(t)$ be the underlying aggregated CTMC for G_i. Then define:

$$\theta_i := \inf\{t \in \mathbb{R}_+ \,:\, I^i_P(t) \in T\} \tag{3.6}$$

Then $\{\theta_i\}_{i=1}^{\infty}$ is a sequence of individual passage-times (Definition 2.2). Recall from Section 2.2, that we have:

$$\mathbb{P}\{\theta_i \leq t\} = \mathbb{P}\{I^i_P(t) \in T\} = \frac{1}{\mathcal{S}(G_i,\, H)} \sum_{Q \in T} \mathbb{E}[N^{G_i}_{H,\, Q}(t)] \tag{3.7}$$

By structural equivalence this is, $\frac{1}{k\mathcal{S}(G_i)} \sum_{Q \in T} \mathbb{E}[N^{G_i}_{H,\, Q}(t)] = \frac{1}{k} \sum_{Q \in T} \mathbb{E}[\bar{N}^{G_i}_{H,\, Q}(t)]$ for some $k > 0$, independent of i. The approximation of Section 2.2 can then be expressed as:

$$\mathbb{P}\{\theta_i \leq t\} \approx \frac{1}{k\mathcal{S}(G_i)} \sum_{Q \in T} v^{G_i}_{H,\, Q}(t) = \frac{1}{k} \sum_{Q \in T} \bar{v}_{H,\, Q}(t) \tag{3.8}$$

since we recall from the last section that the rescaled differential equation approximation is independent of i. The following theorem shows that these quantities become equal in the limit of large population sizes.

Theorem 3.6. *Fix a sequence of structurally equivalent grouped PEPA models, $\{G_i\}_{i=1}^{\infty}$, with all of the above notation. Fix also some component group $H \in \mathcal{G}$ such that $N^{G_i}_{H,\, P}(0) = \mathcal{S}(G_i,\, H)$ for any i and some standard PEPA component, P, and some absorbing set of local target states, $T \subseteq ds(P)$, so that the sequence $\{\theta_i\}_{i=1}^{\infty}$ is a sequence of individual passage-times.*

Fix $S > 0$. Assume that $\mathcal{S}(G_i) \to \infty$ as $i \to \infty$. Then, for all $t \in [0,\, S]$:

$$\mathbb{P}\{\theta_i \leq t\} = \frac{1}{k} \sum_{Q \in T} \mathbb{E}[\bar{N}^{G_i}_{H,\, Q}(t)] \longrightarrow \frac{1}{k} \sum_{Q \in T} \bar{v}_{H,\, Q}(t)$$

as $i \to \infty$.

Proof. Follows from the proof of Theorem 3.5 (Appendix C.5). □

This theorem is illustrated by Figure 8. Note that we would expect the convergence here to happen as fast as in the case of Theorem 3.5 since it also relies only on a convergence of expectations and not a concentration of measure as in the case of Theorem 3.4. Figure 8 supports this claim empirically.

4. Worked Example and Higher-order Moment Global Passage-time Bounds

In Section 2, we introduced a number of schemes for approximating passage-time measures in large Markov models and in Section 3, we showed how these approximations are correct in the limit of large population sizes. In this section, we apply these techniques to a larger, more realistic worked example.

Furthermore, recall that, for global passage-times, we defined two approximation schemes. The first (Section 2.1.1) yielded a deterministic approximation for the passage-time, the accuracy of which requires concentration of the measure around this quantity. However, we also introduced a second scheme (Section 2.1.2), which requires only convergence of expectations, and can thus be expected to be accurate for much smaller population sizes, where there still may be significant variability in the passage-time of interest. The downside is that it yields only approximations to bounds on the cumulative distribution function, as opposed to approximations to the actual function itself. We also use this last section to show, in the context of our worked example, how these approximate bounds on global passage-times can be tightened by considering higher-order moments.

4.1. Two-stage Fetch Client/Server Model

We begin by introducing the model which will form the basis of our worked example. Define the grouped PEPA model, $CS(n, m)$ as below. We have a population of n clients and a population of m servers. The system uses a 2-stage fetch mechanism: a client requests data from the pool of servers; one of the servers receives the request, another server may then later fetch the data for the client. The servers also have some unrelated work to complete. A server in the pool may also fail when performing the actions for the client.

$$\begin{aligned}
\mathit{Client} &\stackrel{def}{=} (\mathit{request}, r_{req}).\mathit{Client_waiting} \\
\mathit{Client_waiting} &\stackrel{def}{=} (\mathit{data}, r_{data}).\mathit{Client_finished} \\
\mathit{Client_finished} &\stackrel{def}{=} \mathit{stop}
\end{aligned}$$

$$\begin{aligned}
\mathit{Server} &\stackrel{def}{=} (\mathit{request}, r_{req}).\mathit{Server_done} + (\mathit{data}, r_{data}).\mathit{Server_done} + (\mathit{break}, r_{break}).\mathit{Server_broken} \\
\mathit{Server_done} &\stackrel{def}{=} (\mathit{work}, r_{work}).\mathit{Server} \\
\mathit{Server_broken} &\stackrel{def}{=} (\mathit{reset}, r_{reset}).\mathit{Server}
\end{aligned}$$

$$CS(n, m) \stackrel{def}{=} \mathbf{Clients}\{\mathit{Client}[n]\} \underset{L}{\bowtie} \mathbf{Servers}\{\mathit{Server}[m]\} \tag{4.1}$$

where $L = \{\mathit{request}, \mathit{data}\}$.

Applying Definition 1.4 yields the system of differential equations in Appendix B.2.1.

4.2. Global Passage-times

In this section, we show how the techniques of Section 2.1 can be used to compute approximations to global passage-times for our worked example. Specifically, we will consider the instance of Definition 2.1 obtained by setting $G = CS(2n, m)$, $\mathbf{c} = (1, 1, 0, 0, 0, 0)$ and $C = n$. We use the ordering given in Equation (4.1) for the vectors $\mathbf{v}(t)$ and $\mathbf{N}(t)$, so this passage-time is for half of the clients to finish.

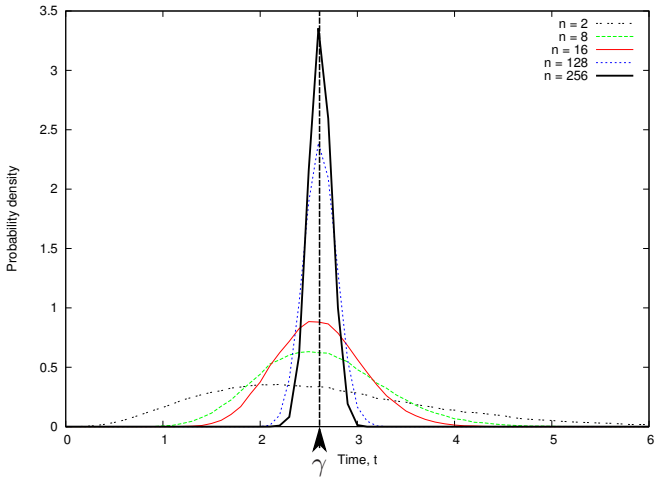


Fig. 9: Probability density functions of passage-time for half of the clients to finish in the model $CS(2n, n)$ compared with the ODE approximation, denoted γ. Rates: $r_{req} = 2.0$, $r_{work} = 2.0$, $r_{break} = 0.5$, $r_{data} = 1.0$ and $r_{reset} = 2.0$.

Figure 9 shows probability density functions computed using traditional methods for this passage-time random variable for different values of n and m, yielding a sequence of structurally equivalent grouped PEPA models with an increasing total component population size. In accordance with Theorem 3.4, the probability density functions are converging towards the point mass at $\gamma := \inf\{t \in \mathbb{R}_+ : \mathbf{c} \cdot \mathbf{v}(t) \leq C\}$.

It is clear that this deterministic approximation in Figure 9 is only really accurate enough to be useful in the cases $n = 128$ and $n = 256$. In case we were interested in, say, $n = 16$, where there is still significant variability in the passage-time, we might try applying the techniques of Section 2.1.2. Doing so yields approximate upper and lower bounds on the cumulative distribution function for the passage-time, illustrated in Figure 10. We have also shown the actual bound approximated by the differential equations. We see that the bounds are fairly tight and furthermore, that the differential equation approximation to them is very accurate. If bounds are good enough metrics, say for example, in the case of validating service level agreement guarantees, this second technique can be more valuable when component populations are only fairly large. In the next section, we show how we can approximate even tighter bounds than those delivered by Markov's inequality.

4.3. Higher-order Moment Approximations to Global Passage-time Bounds

In this section, we show how the techniques of Section 2.1.2 can be extended, by exploiting differential equation approximations to higher-order moments, allowing us to replace the use of Markov's inequality with Chebyshev's inequality, leading to tighter bounds.

We have already introduced the idea of considering the differential equation solutions, $\mathbf{v}(t)$ as approximations to the first moments of the component counting processes of a grouped PEPA model, $\mathbb{E}[\mathbf{N}(t)]$. In fact, it can be shown [11], that the equality $\dot{\mathbb{E}}[\mathbf{N}(t)] = \mathbb{E}[\mathbf{f}(\mathbf{N}(t))]$ holds exactly. The approximation of $\mathbb{E}[\mathbf{N}(t)]$ by $\mathbf{v}(t)$ can then be obtained by approximating $\mathbb{E}[\mathbf{f}(\mathbf{N}(t))]$ by $\mathbf{f}(\mathbb{E}[\mathbf{N}(t)])$. In all but the simplest of cases, $\mathbf{f}(\cdot)$ is non-linear and this is indeed an approximation. As we have seen, however, it often works very well.

Furthermore, it has been shown [11] how similar systems of differential equations can be constructed to approximate higher-order (joint) moments of the component counting stochastic processes, facilitating approximation of, for example, variances. We show how such differential equations may be derived in the

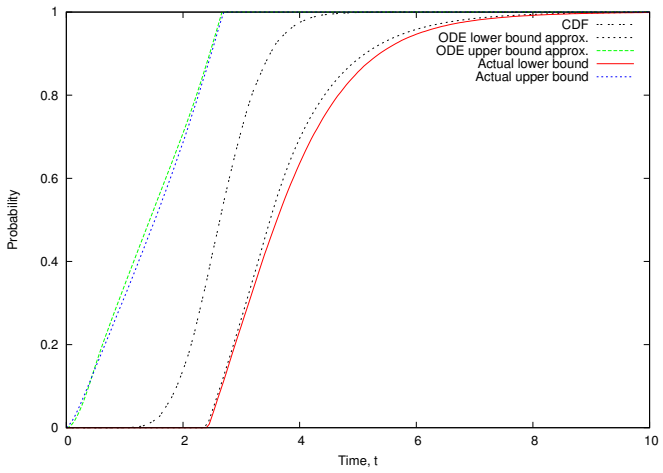


Fig. 10: Cumulative distribution functions of passage-time for half of the clients to finish in the model $CS(16, 8)$ compared with Markov inequality bounds and their ODE approximations. Rates: $r_{req} = 2.0$, $r_{work} = 2.0$, $r_{break} = 0.5$, $r_{data} = 1.0$ and $r_{reset} = 2.0$.

next section, before showing how they can be applied to compute tighter approximate global passage-time bounds in the proceeding section.

4.3.1. Differential equations approximating higher-order moments

In this section, we show how differential equations approximating higher-order moments of $\mathbf{N}(t)$ may be constructed for a grouped PEPA model. The application of fluid-generated higher moments to passage-time approximations (which follows in Section 4.3.2) is a novel contribution.

Differential equations for higher-order moments of component counts can be derived systematically from the underlying PEPA model (Hayden *et al.* [11]). In the example below, we do not repeat this approach and instead derive these equations directly from the state space of the underlying Markov model.

We introduce the general idea using our worked example model, $CS(n, m)$. Since the variance of a component count is generally a quantity much smaller in magnitude than its expectation, an approximation to it tends to introduce more relative numerical error. However, good approximations can be obtained for so-called *split-free* models [11], for which the nature of the ODE approximation remains relatively simple. A split-free model can be defined as one for which rational functions of component counts do not appear in the system of ODEs. Accordingly, the methods of this section and the next will thus only be applicable to split-free models, however this is still a large-class of models, including $CS(n, m)$.

We proceed by considering the Chapman–Kolmogorov equations of the underlying aggregated CTMC. Write $p_{(C,\, C_w,\, C_f,\, S,\, S_d,\, S_b)}(t)$ for the transient probability of being in the aggregate CTMC state of $C \times \mathit{Client}$, $C_w \times \mathit{Client\_waiting}$, $C_f \times \mathit{Client\_finished}$, $S \times \mathit{Server}$, $S_d \times \mathit{Server\_done}$ and $S_b \times \mathit{Server\_broken}$ components

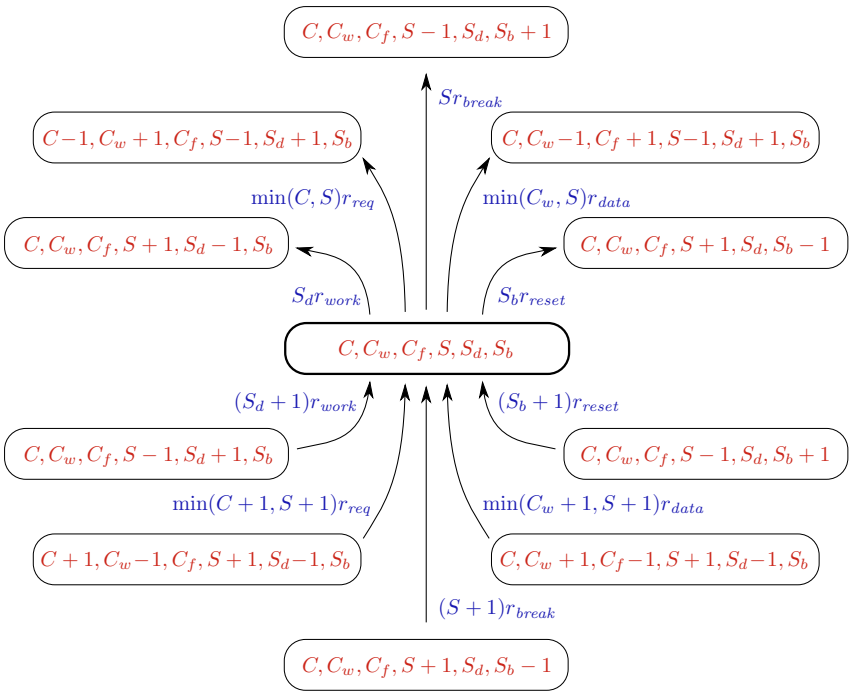


Fig. 11: A central state of the underlying aggregate CTMC of the model $CS(n, m)$

at time t. From Figure 11, we obtain the Chapman–Kolmogorov forward equations:

$$\begin{aligned}
\dot{p}_{(C,\,C_w,\,C_f,\,S,\,S_d,\,S_b)}(t) = {} & r_{req}\min(C+1,\,S+1)\cdot p_{(C+1,\,C_w-1,\,C_f,\,S+1,\,S_d-1,\,S_b)}(t) \\
& + r_{data}\min(C_w+1,\,S+1)\cdot p_{(C,\,C_w+1,\,C_f-1,\,S+1,\,S_d-1,\,S_b)}(t) \\
& + r_{break}(S+1)\cdot p_{(C,\,C_w,\,C_f,\,S+1,\,S_d,\,S_b-1)}(t) \\
& + r_{work}(S_d+1)\cdot p_{(C,\,C_w,\,C_f,\,S-1,\,S_d+1,\,S_b)}(t) \\
& + r_{reset}(S_b+1)\cdot p_{(C,\,C_w,\,C_f,\,S-1,\,S_d,\,S_b+1)}(t) \\
& - (r_{req}\min(C,\,S) + r_{data}\min(C_w,\,S) + r_{break}S + r_{work}S_d \\
& \quad + r_{reset}S_b)\cdot p_{(C,\,C_w,\,C_f,\,S,\,S_d,\,S_b)}(t)
\end{aligned} \tag{4.2}$$

where the first five terms appear only when the subscript of the corresponding $p_{\cdot}(t)$ term is actually in the state space, or we can just fix $p_s(t) := 0$ for all s outside the boundary of the aggregated state space.

We will write $N_C(t)$ for the *Client* counting stochastic process, $N_{S_d}(t)$ for the *Server_done* counting stochastic process and similarly for all other components. Then, for example, if S is the aggregated state space, we have by definition:

$$\dot{\mathbb{E}}[N_C^2(t)] = \sum_{(C,\,C_w,\,C_f,\,S,\,S_d,\,S_b)\in S} \dot{p}_{(C,\,C_w,\,C_f,\,S,\,S_d,\,S_b)}(t)\,C^2 \tag{4.3}$$

We now substitute Equation (4.2) into this. The first term of Equation (4.2) will yield the following contribution to $\dot{\mathbb{E}}[N_C^2(t)]$:

$$\sum_{(C,\,C_w,\,C_f,\,S,\,S_d,\,S_b)\in S} [r_{req}\,C^2\min(C+1,\,S+1)\cdot p_{(C+1,\,C_w-1,\,C_f,\,S+1,\,S_d-1,\,S_b)}(t)] \tag{4.4}$$

Re-indexing the summation carefully gives:

$$\sum_{(C,\, C_w,\, C_f,\, S,\, S_d,\, S_b)\in S} [r_{req}(C-1)^2 \min(C,\, S) \cdot p_{(C,\, C_w,\, C_f,\, S,\, S_d,\, S_b)}(t)] \tag{4.5}$$

Expanding the factor $(C-1)^2$, we can rewrite this as:

$$\sum_{(C,\, C_w,\, C_f,\, S,\, S_d,\, S_b)\in S} [p_{(C,\, C_w,\, C_f,\, S,\, S_d,\, S_b)}(t) \cdot (r_{req} \min(C^3,\, C^2 S) - 2r_{req} \min(C^2,\, CS) + r_{req} \min(C,\, S))] \tag{4.6}$$

which is simply:

$$r_{req}\mathbb{E}[\min(N_C^3(t),\, N_C^2(t)N_S(t))] - 2r_{req}\mathbb{E}[\min(N_C^2(t),\, N_C(t)N_S(t))] + r_{req}\mathbb{E}[\min(N_C(t),\, N_S(t))] \tag{4.7}$$

Note that the first term of the above is cancelled by the contribution of the first negative term of Equation (4.2). The contributions of the remaining four positive terms of Equation (4.2) to $\dot{\mathbb{E}}[N_C^2(t)]$ all cancel with their corresponding negative counterparts, and we get the following:

$$\dot{\mathbb{E}}[N_C^2(t)] = -2r_{req}\mathbb{E}[\min(N_C^2(t),\, N_C(t)N_S(t))] + r_{req}\mathbb{E}[\min(N_C(t),\, N_S(t))] \tag{4.8}$$

Considering all of the remaining (joint) first- and second-order moments of the component counting processes in a similar fashion yields a system of 27 such equations. These cannot be solved exactly on their own since they involve expectations of non-linear expressions. However, we can apply an approximation to their right-hand sides in the same spirit as for the first order case discussed above to obtain a system of approximating coupled differential equations.

We will write $v_{C^2}(t)$ for the approximation to the second moment of the *Client* counting process and $v_{C\cdot S_d}(t)$ for the approximation to the joint moment of the *Client* and *Server_done* counting processes, and again, similarly for all other first and second order moment approximations. Explicitly, we intend that, $v_{C\cdot S_d}(t) \approx \mathbb{E}[N_C(t)N_{S_d}(t)]$ and so on. Specifically, we make the approximation $\mathbb{E}[\min(X,\, Y)] \approx \min(\mathbb{E}[X],\, \mathbb{E}[Y])$ in Equation (4.8), to obtain our actual moment approximation:

$$\dot{v}_{C^2}(t) = -2r_{req} \min(v_{C^2(t)},\, v_{C\cdot S}(t)) + r_{req} \min(v_C(t),\, v_S(t)) \tag{4.9}$$

and similarly for all other first- and second-order moments, yielding a complete system of 27 coupled ODEs, which can be efficiently solved uniquely for the moment approximations. This system is given in its entirety in Appendices B.2.1 and B.2.2.

Similar considerations can generate systems of ODEs approximating arbitrary orders of joint moments, as described in detail in [11], where it is clear that the generation process could be automated with little difficulty.

4.3.2. Global passage-time bounds with Chebyshev's inequality

In this section, we show how we can use the approximation to higher-order moments just introduced combined with Chebyshev's inequality to tighten the approximate bounds of Section 2.1.2.

If X is an arbitrary random variable, $t > 0$, $q \neq 0$, Chebyshev's inequality says:

$$\mathbb{P}\{|X - \mathbb{E}[X]| \geq t\} \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^q]}{t^q} \tag{4.10}$$

In this paper we consider only the case $q = 2$, however it may be possible to obtain tighter passage-time bounds by considering higher integer values of q, at the expense of a larger system of differential equations.

In the case $q = 2$, it is a straightforward application of the Cauchy–Schwarz inequality to derive a one-sided improvement of this, often known as the Chebyshev–Cantelli inequality, which says that for $t > 0$:

$$\mathbb{P}\{X - \mathbb{E}[X] \geq t\} \leq \frac{\mathrm{Var}[X]}{\mathrm{Var}[X] + t^2} \tag{4.11}$$

and symmetrically:

$$\mathbb{P}\{\mathbb{E}[X] - X \geq t\} \leq \frac{\mathrm{Var}[X]}{\mathrm{Var}[X] + t^2} \tag{4.12}$$

To illustrate how we might use these inequalities to improve on the global passage-time bounds of Section 2.1.2, which were obtained using Markov's inequality, consider again the model $CS(2n, m)$ and the global passage-time defined earlier for half (or n) of the clients to finish. Denote this random variable, σ, then we have:

$$\begin{aligned} \mathbb{P}\{\sigma \leq t\} &= \mathbb{P}\{N_{C_f}(t) \geq n\} \\ &= \mathbb{P}\{N_{C_f}(t) - \mathbb{E}[N_{C_f}(t)] \geq n - \mathbb{E}[N_{C_f}(t)]\} \end{aligned} \tag{4.13}$$

and:

$$\mathbb{P}\{\sigma \leq t\} \geq 1 - \mathbb{P}\{\mathbb{E}[N_{C_f}(t)] - N_{C_f}(t) \geq \mathbb{E}[N_{C_f}(t)] - n\} \tag{4.14}$$

Applying the Chebyshev–Cantelli inequality to the probability on the right-hand side in each case then yields the following bounds on the cumulative distribution function of σ:

$$\mathbb{P}\{\sigma \leq t\} \leq \frac{\mathrm{Var}[N_{C_f}(t)]}{\mathrm{Var}[N_{C_f}(t)] + (\mathbb{E}[N_{C_f}(t)] - n)^2} \tag{4.15}$$

and:

$$\mathbb{P}\{\sigma \leq t\} \geq 1 - \frac{\mathrm{Var}[N_{C_f}(t)]}{\mathrm{Var}[N_{C_f}(t)] + (\mathbb{E}[N_{C_f}(t)] - n)^2} \tag{4.16}$$

where the first is valid when $n - \mathbb{E}[N_{C_f}(t)] > 0$, and the second when $\mathbb{E}[N_{C_f}(t)] - n > 0$. We may then apply the differential equation approximations to first- and second-order moments defined in the previous section to yield approximations to these bounds.

Figure 12 shows how these bounds substantially improve on those obtained using Markov's inequality (Figure 10). However, we do notice that the differential equation approximation deviates more substantially from the actual bound in the Chebyshev case. This is to be expected due to the higher relative error in the variance approximations.

4.4. Individual Passage-times

Finally, in this section, we show how the techniques of Section 2.2 can be used to compute approximations to individual passage-times for our worked example. Specifically, we will consider the instance of Definition 2.2 obtained by setting $G = CS(2n, m)$, $H = \mathbf{Clients}$ and $T = \{Client\_finished\}$. That is, the time for a given client to finish.

Figure 13 shows cumulative distribution functions computed using traditional methods for this passage-time random variable for different values of n and m, yielding a sequence of structurally equivalent grouped PEPA models with an increasing total component population size. In accordance with Theorem 3.6, the cumulative distribution functions are converging towards the differential equation approximation.

5. Conclusion

Passage-time measures in Markov chains are extremely useful for expressing probabilistic durations in real-world applications. Until now these calculations were limited to explicit state models and were limited by the size of system being analysed.

In this paper, we have applied recent developments in fluid analysis of large Markov chains to allow us to approximate passage-time distributions as well. We introduced the notion of global and individual passage times as being useful quantities in the context of massively parallel systems.

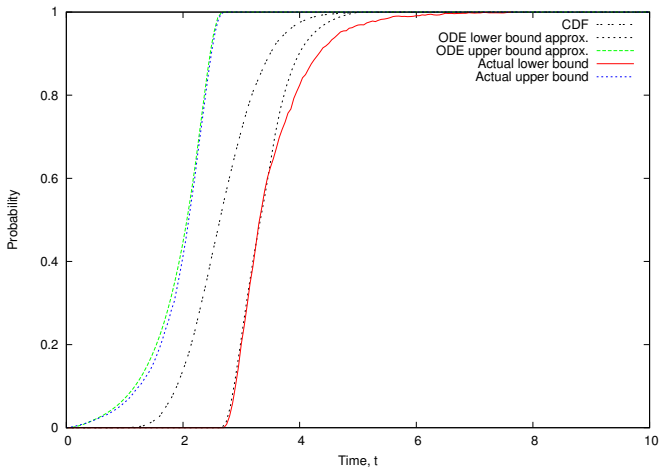


Fig. 12: Cumulative distribution functions of passage-time for half of the clients to finish in the model $CS(16, 8)$ compared with Chebyshev inequality bounds and their ODE approximations. Rates: $r_{req} = 2.0$, $r_{work} = 2.0$, $r_{break} = 0.5$, $r_{data} = 1.0$ and $r_{reset} = 2.0$.

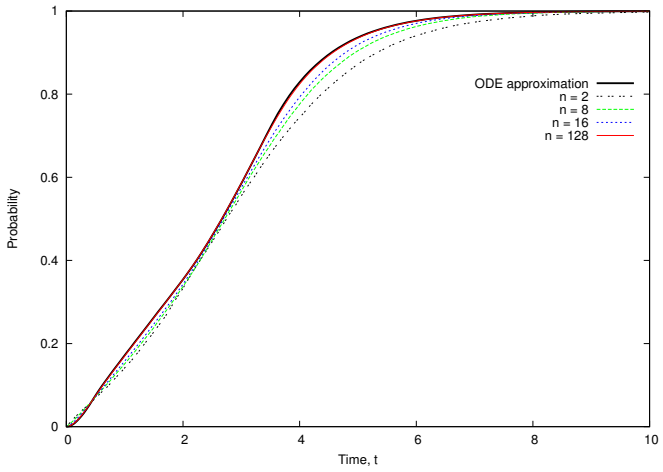


Fig. 13: Cumulative distribution functions (CDFs) of passage-time for one client to finish in the model $CS(2n, n)$ compared with the ODE approximation to the CDF. Rates: $r_{req} = 2.0$, $r_{work} = 2.0$, $r_{break} = 0.5$, $r_{data} = 1.0$ and $r_{reset} = 2.0$.

Firstly, we showed a limiting result for global passage times (Theorem 3.4), that for sequences of structurally similar Markov chains, the distribution of the actual passage time tends towards the deterministic approximation obtained via fluid analysis.

Secondly, also for global passage times and using two standard probabilistic inequalities, we have shown that fluid techniques can establish approximations for both upper and lower bounds of the cumulative distribution function of system-wide passage times in the Markov chain (Theorem 3.5).

Finally, for individual passage times, we proved that a set of fluid approximations can be used to generate the entire cumulative distribution function of the required passage time (Theorem 3.6).

We demonstrated these techniques on example Markov chains of the order of 2^{100} states. For the global passage-time analysis, where the scale of the model is sufficiently large, we observe that the deterministic approximation to the passage-time is reasonably accurate. In cases where there is still significant variability in the passage-time distribution, we can obtain accurate fluid approximations to CDF bounds. These bounds can, in some cases, be fairly loose, and we have shown that the situation can be improved somewhat by considering higher-order moment approximations using the Chebyshev inequality. It is important to realise that a CDF lower bound for a passage time is conservatively sufficient for verifying SLAs specified in terms of passage-time quantiles. Thus the bounds are directly useful, even when (as with the Markov inequality) the bound itself can, in some cases, be fairly loose.

In the individual passage time case, the fluid approximation converges relatively quickly to the CDF, making it particularly useful from an engineering perspective.

Future work includes trying to expand the type of passage time structure that can be calculated or bounded in this way. For the limiting results of Theorem 3.4 and Theorem 3.6, we wish to investigate the rate of convergence. Potentially this could provide quantitative bounds on distribution of passage-time random variables.

References

[1] J. T. Bradley, N. J. Dingle, S. T. Gilmore, and W. J. Knottenbelt, "Derivation of passage-time densities in PEPA models using ipc: the Imperial PEPA Compiler," in *MASCOTS'03, Proceedings of the 11th IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunications Systems* (G. Kotsis, ed.), (University of Central Florida), pp. 344–351, IEEE Computer Society Press, October 2003.

[2] J. T. Bradley, R. Hayden, W. J. Knottenbelt, and T. Suto, "Extracting Response Times from Fluid Analysis of Performance Models," in *SIPEW'08, SPEC International Performance Evaluation Workshop, Darmstadt, 27-28 June 2008*, vol. 5119 of *Lecture Notes in Computer Science*, pp. 29–43, May 2008.

[3] J. Hillston, *A Compositional Approach to Performance Modelling*. Cambridge University Press, 1996.

[4] J. Hillston, "Process algebras for quantitative analysis," in *Proceedings of the 20th Annual IEEE Symposium on Logic in Computer Science (LICS' 05)*, (Chicago), pp. 239–248, IEEE Computer Society Press, June 2005.

[5] H. Bowman, J. Bryans, and J. Derrick, "Analysis of a multimedia stream using stochastic process algebra," in *Sixth International Workshop on Process Algebras and Performance Modelling* (C. Priami, ed.), (Nice), pp. 51–69, September 1998.

[6] J. Forneau, L. Kloul, and F. Valois, "Performance modelling of hierarchical cellular networks using PEPA," *Performance Evaluation*, vol. 50, pp. 83–99, Nov. 2002.

[7] N. Thomas, J. T. Bradley, and W. J. Knottenbelt, "Stochastic analysis of scheduling strategies in a GRID-based resource model," *IEE Software Engineering*, vol. 151, pp. 232–239, September 2004.

[8] D. R. W. Holton, "A PEPA specification of an industrial production cell," in *Process Algebra and Performance Modelling Workshop* (S. Gilmore and J. Hillston, eds.), vol. 38(7) of *Special Issue: The Computer Journal*, pp. 542–551, CEPIS, Edinburgh, June 1995.

[9] J. Bradley, N. Dingle, S. Gilmore, and W. Knottenbelt, "Derivation of passage-time densities in PEPA models using IPC: The Imperial PEPA Compiler," in *Proceedings of the 11th IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunications Systems* (G. Kotsis, ed.), (University of Central Florida), pp. 344–351, IEEE Computer Society Press, Oct. 2003.

[10] A. Duguid, "Coping with the parallelism of BitTorrent: Conversion of PEPA to ODEs in dealing with state space explosion," in *Formal Modeling and Analysis of Timed Systems, 4th International Conference, FORMATS 2006, Paris, France, September 25-27, 2006, Proceedings* (E. Asarin and P. Bouyer, eds.), vol. 4202 of *Lecture Notes in Computer Science*, pp. 156–170, Springer, 2006.

[11] R. Hayden and J. Bradley, "A fluid analysis framework for a Markovian process algebra," *Theoretical Computer Science*, 2008. (submitted).

[12] R. A. Hayden and J. T. Bradley, "Fluid semantics for passive stochastic process algebra cooperation," in *VALUETOOLS'08, Third International Conference on Performance Evaluation Methodologies and Tools*, (Athens), 2008.

[13] J. Hillston, "Fluid flow approximation of PEPA models," in *Proceedings of the Second International Conference on the Quantitative Evaluation of Systems*, (Torino, Italy), pp. 33–43, IEEE Computer Society Press, Sept. 2005.

[14] J. T. Bradley, S. T. Gilmore, and J. Hillston, "Analysing distributed internet worm attacks using continuous state-space approximation of process algebra models," *Journal of Computer and System Sciences*, vol. 74, pp. 1013–1032, September 2008.

[15] N. Geisweiller, J. Hillston, and M. Stenico, "Relating continuous and discrete PEPA models of signalling pathways," *Theoretical Computer Science*, vol. 404, pp. 97–111, November 2008.

[16] L. Bortolussi and A. Policriti, "Stochastic concurrent constraint programming and differential equations," in *QAPL'07,*

5th Workshop on Quantitative Aspects of Programming Languages, vol. 190 of *Electronic Notes in Theoretical Computer Science*, pp. 27–42, September 2007.

[17] L. Cardelli, "From processes to ODEs by Chemistry," in *TCS 2008, Fifth IFIP International Conference on Theoretical Computer Science*, (Milan), Springer, 2008.

[18] J. Júlvez, E. Jiménez, L. Recalde, and M. Silva, "On observability in timed continuous petri net systems," in *QEST'04, 1st International Conference on Quantitative Evaluation of Systems*, vol. 266, pp. 60–69, IEEE, September 2004.

[19] N. J. Dingle, P. G. Harrison, and W. J. Knottenbelt, "Uniformization and hypergraph partitioning for the distributed computation of response time densities in very large Markov models," *Journal of Parallel and Distributed Computing*, vol. 64, pp. 908–920, August 2004.

[20] O. Kallenberg, *Foundations of Modern Probability*. Springer, 2002.

[21] R. W. R. Darling and J. R. Norris, "Differential equation approximations for Markov chains," *Probability surveys*, vol. 5, p. 37, 2008.

[22] R. W. R. Darling, "Fluid limits of pure jump Markov processes: a practical guide," tech. rep., National Security Agency, 2002.

[23] T. Kurtz and S. Ethier, *Markov Processes Characterisation and Convergence*. Wiley, 1986.

A. PEPA Functions: Formal Definitions

A.1. Apparent Rate

We define the notion of *apparent rate* as the externally observed rate of activities of a particular type. For a given action type $\alpha \in \mathcal{A}$, it is thus calculated by summing the rates of all enabled activities of this type:

$$r_\alpha(P) := \sum_{P \xrightarrow{(\alpha,\,\lambda)}} \lambda$$

where $\lambda \in \mathbb{R}_+$.

Apparent rate can also be defined equivalently in a recursive manner over the PEPA grammar as follows:

$$\begin{aligned} r_\alpha((\beta,\,\lambda).P) &:= \begin{cases} \lambda & \text{if } \beta = \alpha \\ 0 & \text{if } \beta \neq \alpha \end{cases} \\ r_\alpha(P + Q) &:= r_\alpha(P) + r_\alpha(Q) \\ r_\alpha(P/L) &:= \begin{cases} r_\alpha(P) & \text{if } \alpha \notin L \\ 0 & \text{if } \alpha \in L \end{cases} \\ r_\alpha(P \underset{L}{\bowtie} Q) &:= \begin{cases} \min(r_\alpha(P),\, r_\alpha(Q)) & \text{if } \alpha \in L \\ r_\alpha(P) + r_\alpha(Q) & \text{if } \alpha \notin L \end{cases} \end{aligned} \tag{A.1}$$

A.2. PEPA Flattening Function

Definition A.1 (Model flattening function)**.** *For any grouped PEPA model G, the corresponding standard PEPA model, $\mathcal{F}(G)$, can be recovered from the grouped model. $\mathcal{F}(\cdot)$ is defined as:*

$$\begin{aligned} \mathcal{F}(M_1 \underset{L}{\bowtie} M_2) &:= \mathcal{F}(M_1) \underset{L}{\bowtie} \mathcal{F}(M_2) \\ \mathcal{F}(Y\{D\}) &:= \mathcal{F}'(D) \end{aligned}$$

where for component groups:

$$\begin{aligned} \mathcal{F}'(D_1 \parallel D_2) &:= \mathcal{F}'(D_1) \parallel \mathcal{F}'(D_2) \\ \mathcal{F}'(P) &:= P \end{aligned}$$

The following theorem proved in [11] states formally the intention that a grouped PEPA model behaves exactly as the corresponding standard PEPA model. For the purposes of this work, this can be taken as the definition of the operational semantics for grouped PEPA models.

Theorem A.2. *Let G be a grouped PEPA model. Then for all $\alpha \in \mathcal{A}$, transitions $G \xrightarrow{(\alpha, r)} G'$ are in one-to-one correspondence with transitions $\mathcal{F}(G) \xrightarrow{(\alpha, r)} \mathcal{F}(G')$.*

Furthermore, the apparent rate of a grouped PEPA model G is defined exactly in terms of the corresponding standard PEPA model, i.e. $r_\alpha(G) := r_\alpha(\mathcal{F}(G))$.

A.3. PEPA Group Functions

Definition A.3 (Set of component group labels)**.** *For any grouped PEPA model G, its set of component group labels is $\mathcal{G}(G)$ where $\mathcal{G}(M_1 \underset{L}{\bowtie} M_2) := \mathcal{G}(M_1) \cup \mathcal{G}(M_2)$ and $\mathcal{G}(Y\{D\}) := Y$.*

Definition A.4 (Standard PEPA derivative states in a component group)**.** *For any grouped PEPA model G, the set of standard PEPA component derivative states in a given component group with label $H \in \mathcal{G}(G)$ is $\mathcal{B}(G, H)$ where:*

$$\mathcal{B}(M_1 \underset{L}{\bowtie} M_2, H) := \begin{cases} \mathcal{B}(M_1, H) & \text{if } H \in \mathcal{G}(M_1) \\ \mathcal{B}(M_2, H) & \text{if } H \in \mathcal{G}(M_2) \end{cases}$$

and $\mathcal{B}(H\{D\}, H) := \mathcal{B}'(D)$. For component groups, $\mathcal{B}'(D_1 \parallel D_2) := \mathcal{B}'(D_1) \cup \mathcal{B}'(D_2)$ and $\mathcal{B}'(P) := ds(P)$.

Furthermore define $\mathcal{B}(G)$ to be the subset of $\mathcal{G}(G) \times \bigcup_{H \in \mathcal{G}(G)} \mathcal{B}(G, H)$ such that $(H, P) \in \mathcal{B}(G)$ if and only if $H \in \mathcal{G}(G)$ and $P \in \mathcal{B}(G, H)$. That is, there is exactly one element of $\mathcal{B}(G)$ for every standard PEPA component and group in which it occurs in the model. This allows us to specify the standard PEPA components of a particular type occurring in a given component group.

We write also $\mathcal{S}(G, H)$ for the size of the component group $H \in \mathcal{G}(G)$, that is, the number of parallel components in the group. So, for example, $\mathcal{S}(PR_G(n, m), \mathbf{Processors}) = n$. We then write $\mathcal{S}(G) := \sum_{H \in \mathcal{G}(G)} \mathcal{S}(G, H)$ for the total component population of the model G.

Lastly, for fairly technical reasons concerned with the proof of Lemma 1.5, we will need to know for an action type $\alpha \in \mathcal{A}$, the *structural depth* of a grouped PEPA model, G. This is simply the largest number of cooperations involving α, whose immediate effect can be seen by a standard PEPA component enabling an α-action within some component group.

Definition A.5 (Structural depth)**.** *For any grouped PEPA model G and action type $\alpha \in \mathcal{A}$, the structural depth of G with respect to α is $\mathcal{D}_\alpha(G)$ where $\mathcal{D}_\alpha(\cdot)$ is defined as:*

$$\mathcal{D}_\alpha(M_1 \underset{L}{\bowtie} M_2) := \begin{cases} 1 + \max\{\mathcal{D}_\alpha(M_1), \mathcal{D}_\alpha(M_2)\} & \text{if } \alpha \in L \\ \max\{\mathcal{D}_\alpha(M_1), \mathcal{D}_\alpha(M_2)\} & \text{if } \alpha \notin L \end{cases}$$

$$\mathcal{D}_\alpha(Y\{D\}) := 0$$

B. Systems of Equations

B.1. Differential Equations for $PR'_G(n, m)$

We have adopted the shorthand $v_{P_0}(t)$ for $v_{\mathbf{Processors}, Processor_0}(t)$ and similarly for the other quantities.

$$\begin{aligned}
\dot{v}_{P_0}(t) &= -\frac{v_{P_0}(t)}{v_{P_0}(t) + v_{P'_0}(t)} \min(r_1(v_{P_0}(t) + v_{P'_0}(t)),\, r_2 v_{R_0}(t)) \\
\dot{v}_{P_1}(t) &= -q_1 v_{P_1}(t) + \frac{v_{P_0}(t)}{v_{P_0}(t) + v_{P'_0}(t)} \min(r_1(v_{P_0}(t) + v_{P'_0}(t)),\, r_2 v_{R_0}(t)) \\
\dot{v}_{P'_0}(t) &= -\frac{v_{P'_0}(t)}{v_{P_0}(t) + v_{P'_0}(t)} \min(r_1(v_{P_0}(t) + v_{P'_0}(t)),\, r_2 v_{R_0}(t)) + q_1(v_{P_1}(t) + v_{P'_1}(t)) \\
\dot{v}_{P'_1}(t) &= -q_1 v_{P'_1}(t) + \frac{v_{P'_0}(t)}{v_{P_0}(t) + v_{P'_0}(t)} \min(r_1(v_{P_0}(t) + v_{P'_0}(t)),\, r_2 v_{R_0}(t)) \\
\dot{v}_{R_0}(t) &= -\min(r_1(v_{P_0}(t) + v_{P'_0}(t)),\, r_2 v_{R_0}(t)) + q_2 v_{R_1}(t) \\
\dot{v}_{R_1}(t) &= -q_2 v_{R_1}(t) + \min(r_1(v_{P_0}(t) + v_{P'_0}(t)),\, r_2 v_{R_0}(t))
\end{aligned} \tag{B.1}$$

B.2. Differential Equations for CS(n, m)

B.2.1. First Order Moments

We have adopted the shorthand $v_C(t)$ for $v_{\mathbf{Clients},\,Client}(t)$ and similarly for the other quantities.

$$
\begin{aligned}
\dot{v}_C(t) &= -r_{req}\min(v_C(t),\, v_S(t)) \\
\dot{v}_{C_w}(t) &= r_{req}\min(v_C(t),\, v_S(t)) - r_{data}\min(v_{C_w}(t),\, v_S(t)) \\
\dot{v}_{C_f}(t) &= r_{data}\min(v_{C_w}(t),\, v_S(t)) \\
\dot{v}_S(t) &= -r_{req}\min(v_C(t),\, v_S(t)) - r_{data}\min(v_{C_w}(t),\, v_S(t)) + r_{work}v_{S_d}(t) - r_{break}v_S(t) + r_{reset}v_{S_b}(t) \\
\dot{v}_{S_d}(t) &= r_{req}\min(v_C(t),\, v_S(t)) + r_{data}\min(v_{C_w}(t),\, v_S(t)) - r_{work}v_{S_d}(t) \\
\dot{v}_{S_b}(t) &= r_{break}v_S(t) - r_{reset}v_{S_b}(t)
\end{aligned}
\tag{B.2}
$$

B.2.2. Second Order Moments

As in Section 4.3.1, we have adopted the shorthand $v_{C^2}(t)$ for the approximation to the second moment of the *Client* counting process and $v_{C\cdot S_d}(t)$ for the approximation to the joint moment of the *Client* and *Server_done* counting processes, and again, similarly for all other second order moment differential equation components.

$$
\begin{aligned}
\dot{v}_{C^2}(t) &= r_{req}(-\min(v_{C^2}(t),\, v_{C\cdot S}(t)) - \min(v_{C^2}(t),\, v_{C\cdot S}(t)) + \min(v_C(t),\, v_S(t))) \\
\dot{v}_{C\cdot C_w}(t) &= r_{req}(-\min(v_{C\cdot C_w}(t),\, v_{C_w\cdot S}(t)) + \min(v_{C^2}(t),\, v_{C\cdot S}(t)) - \min(v_C(t),\, v_S(t))) \\
&\quad - r_{data}\min(v_{C\cdot C_w}(t),\, v_{C\cdot S}(t)) \\
\dot{v}_{C\cdot C_f}(t) &= -r_{req}\min(v_{C\cdot C_f}(t),\, v_{C_f\cdot S}(t)) + r_{data}\min(v_{C\cdot C_w}(t),\, v_{C\cdot S}(t)) \\
\dot{v}_{C\cdot S}(t) &= r_{req}(-\min(v_{C\cdot S}(t),\, v_{S^2}(t)) - \min(v_{C^2}(t),\, v_{C\cdot S}(t)) + \min(v_C(t),\, v_S(t))) \\
&\quad - r_{data}\min(v_{C\cdot C_w}(t),\, v_{C\cdot S}(t)) + r_{work}v_{C\cdot S_d}(t) - r_{break}v_{C\cdot S}(t) + r_{reset}v_{C\cdot S_b}(t) \\
\dot{v}_{C\cdot S_d}(t) &= r_{req}(-\min(v_{C\cdot S_d}(t),\, v_{S\cdot S_d}(t)) + \min(v_{C^2}(t),\, v_{C\cdot S}(t)) - \min(v_C(t),\, v_S(t))) \\
&\quad + r_{data}\min(v_{C\cdot C_w}(t),\, v_{C\cdot S}(t)) - r_{work}v_{C\cdot S_d}(t) \\
\dot{v}_{C\cdot S_b}(t) &= -r_{req}\min(v_{C\cdot S_b}(t),\, v_{S\cdot S_b}(t)) + r_{break}v_{C\cdot S}(t) - r_{reset}v_{C\cdot S_b}(t) \\
\dot{v}_{C_w{}^2}(t) &= r_{req}(\min(v_{C\cdot C_w}(t),\, v_{C_w\cdot S}(t)) + \min(v_{C\cdot C_w}(t),\, v_{C_w\cdot S}(t)) + \min(v_C(t),\, v_S(t))) \\
&\quad + r_{data}(-\min(v_{C_w{}^2}(t),\, v_{C_w\cdot S}(t)) - \min(v_{C_w{}^2}(t),\, v_{C_w\cdot S}(t)) + \min(v_{C_w}(t),\, v_S(t))) \\
\dot{v}_{C_w\cdot C_f}(t) &= r_{req}\min(v_{C\cdot C_f}(t),\, v_{C_f\cdot S}(t)) \\
&\quad + r_{data}(-\min(v_{C_w\cdot C_f}(t),\, v_{C_f\cdot S}(t)) + \min(v_{C_w{}^2}(t),\, v_{C_w\cdot S}(t)) - \min(v_{C_w}(t),\, v_S(t))) \\
\dot{v}_{C_w\cdot S}(t) &= r_{req}(\min(v_{C\cdot S}(t),\, v_{S^2}(t)) - \min(v_{C\cdot C_w}(t),\, v_{C_w\cdot S}(t)) - \min(v_C(t),\, v_S(t))) \\
&\quad + r_{data}(-\min(v_{C_w\cdot S}(t),\, v_{S^2}(t)) - \min(v_{C_w{}^2}(t),\, v_{C_w\cdot S}(t)) + \min(v_{C_w}(t),\, v_S(t))) \\
&\quad + r_{work}v_{C_w\cdot S_d}(t) - r_{break}v_{C_w\cdot S}(t) + r_{reset}v_{C_w\cdot S_b}(t) \\
\dot{v}_{C_w\cdot S_d}(t) &= r_{req}(\min(v_{C\cdot S_d}(t),\, v_{S\cdot S_d}(t)) + \min(v_{C\cdot C_w}(t),\, v_{C_w\cdot S}(t)) + \min(v_C(t),\, v_S(t))) \\
&\quad + r_{data}(-\min(v_{C_w\cdot S_d}(t),\, v_{S\cdot S_d}(t)) + \min(v_{C_w{}^2}(t),\, v_{C_w\cdot S}(t)) - \min(v_{C_w}(t),\, v_S(t))) \\
&\quad - r_{work}v_{C_w\cdot S_d}(t) \\
\dot{v}_{C_w\cdot S_b}(t) &= r_{req}\min(v_{C\cdot S_b}(t),\, v_{S\cdot S_b}(t)) - r_{data}\min(v_{C_w\cdot S_b}(t),\, v_{S\cdot S_b}(t)) + r_{break}v_{C_w\cdot S}(t) - r_{reset}v_{C_w\cdot S_b}(t) \\
\dot{v}_{C_f{}^2}(t) &= r_{data}(\min(v_{C_w\cdot C_f}(t),\, v_{C_f\cdot S}(t)) + \min(v_{C_w\cdot C_f}(t),\, v_{C_f\cdot S}(t)) + \min(v_{C_w}(t),\, v_S(t)))
\end{aligned}
$$

$$
\begin{aligned}
\dot{v}_{C_f\cdot S}(t) = &-r_{req}\min(v_{C\cdot C_f}(t),\, v_{C_f\cdot S}(t))\\
&+r_{data}(\min(v_{C_w\cdot S}(t),\, v_{S^2}(t)) - \min(v_{C_w\cdot C_f}(t),\, v_{C_f\cdot S}(t)) - \min(v_{C_w}(t),\, v_S(t))) + r_{work}v_{C_f\cdot S_d}(t)\\
&-r_{break}v_{C_f\cdot S}(t) + r_{reset}v_{C_f\cdot S_b}(t)\\
\dot{v}_{C_f\cdot S_d}(t) = &\, r_{req}\min(v_{C\cdot C_f}(t),\, v_{C_f\cdot S}(t))\\
&+r_{data}(\min(v_{C_w\cdot S_d}(t),\, v_{S\cdot S_d}(t)) + \min(v_{C_w\cdot C_f}(t),\, v_{C_f\cdot S}(t)) + \min(v_{C_w}(t),\, v_S(t))) - r_{work}v_{C_f\cdot S_d}(t)\\
\dot{v}_{C_f\cdot S_b}(t) = &\, r_{data}\min(v_{C_w\cdot S_b}(t),\, v_{S\cdot S_b}(t)) + r_{break}v_{C_f\cdot S}(t) - r_{reset}v_{C_f\cdot S_b}(t)\\
\dot{v}_{S^2}(t) = &\, r_{req}(-\min(v_{C\cdot S}(t),\, v_{S^2}(t)) - \min(v_{C\cdot S}(t),\, v_{S^2}(t)) + \min(v_C(t),\, v_S(t)))\\
&+r_{data}(-\min(v_{C_w\cdot S}(t),\, v_{S^2}(t)) - \min(v_{C_w\cdot S}(t),\, v_{S^2}(t)) + \min(v_{C_w}(t),\, v_S(t)))\\
&+r_{work}(v_{S\cdot S_d}(t) + v_{S\cdot S_d}(t) + v_{S_d}(t)) + r_{break}(-v_{S^2}(t) - v_{S^2}(t) + v_S(t))\\
&+r_{reset}(v_{S\cdot S_b}(t) + v_{S\cdot S_b}(t) + v_{S_b}(t))\\
\dot{v}_{S\cdot S_d}(t) = &\, r_{req}(-\min(v_{C\cdot S_d}(t),\, v_{S\cdot S_d}(t)) + \min(v_{C\cdot S}(t),\, v_{S^2}(t)) - \min(v_C(t),\, v_S(t)))\\
&+r_{data}(-\min(v_{C_w\cdot S_d}(t),\, v_{S\cdot S_d}(t)) + \min(v_{C_w\cdot S}(t),\, v_{S^2}(t)) - \min(v_{C_w}(t),\, v_S(t)))\\
&+r_{work}(v_{S_d{}^2}(t) - v_{S\cdot S_d}(t) - v_{S_d}(t)) - r_{break}v_{S\cdot S_d}(t) + r_{reset}v_{S_d\cdot S_b}(t)\\
\dot{v}_{S\cdot S_b}(t) = &-r_{req}\min(v_{C\cdot S_b}(t),\, v_{S\cdot S_b}(t)) - r_{data}\min(v_{C_w\cdot S_b}(t),\, v_{S\cdot S_b}(t)) + r_{work}v_{S_d\cdot S_b}(t)\\
&+r_{break}(-v_{S\cdot S_b}(t) + v_{S^2}(t) - v_S(t)) + r_{reset}(v_{S_b{}^2}(t) - v_{S\cdot S_b}(t) - v_{S_b}(t))\\
\dot{v}_{S_d{}^2}(t) = &\, r_{req}(\min(v_{C\cdot S_d}(t),\, v_{S\cdot S_d}(t)) + \min(v_{C\cdot S_d}(t),\, v_{S\cdot S_d}(t)) + \min(v_C(t),\, v_S(t)))\\
&+r_{data}(\min(v_{C_w\cdot S_d}(t),\, v_{S\cdot S_d}(t)) + \min(v_{C_w\cdot S_d}(t),\, v_{S\cdot S_d}(t)) + \min(v_{C_w}(t),\, v_S(t)))\\
&+r_{work}(-v_{S_d{}^2}(t) - v_{S_d{}^2}(t) + v_{S_d}(t))\\
\dot{v}_{S_d\cdot S_b}(t) = &\, r_{req}\min(v_{C\cdot S_b}(t),\, v_{S\cdot S_b}(t)) + r_{data}\min(v_{C_w\cdot S_b}(t),\, v_{S\cdot S_b}(t)) - r_{work}v_{S_d\cdot S_b}(t)\\
&+r_{break}v_{S\cdot S_d}(t) - r_{reset}v_{S_d\cdot S_b}(t)\\
\dot{v}_{S_b{}^2}(t) = &\, r_{break}(v_{S\cdot S_b}(t) + v_{S\cdot S_b}(t) + v_S(t)) + r_{reset}(-v_{S_b{}^2}(t) - v_{S_b{}^2}(t) + v_{S_b}(t))
\end{aligned}
\tag{B.3}
$$

C. Proofs and Lemmas

C.1. Proof of Lemma 1.5

Proof. Writing the system of differential equations associated to a grouped PEPA model, G, as $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$. We see from Definition 1.4 that for $1 \leq k \leq \mathcal{N}(G)$ (corresponding to some $(H,\, P) \in \mathcal{B}(G)$) and $\mathbf{v} \in \mathbb{R}_+^{\mathcal{N}(G)}$:

$$
f_k(\mathbf{v}) = \sum_{\alpha_i \in \mathcal{A}} \left(\sum_{Q_j \in \mathcal{B}(G,\, H)} p_{\alpha_i}(Q_j,\, P)\, \mathcal{R}_{\alpha_i}(G,\, \mathbf{v},\, H,\, Q_j) \right) - \mathcal{R}_{\alpha_i}(G,\, \mathbf{v},\, H,\, P)
$$

For arbitrary $\alpha \in \mathcal{A}$, we focus now on a term, $\mathcal{R}_\alpha(G,\, \mathbf{v},\, H,\, Q)$ for $Q \in \mathcal{B}(G,\, H)$. It is an easy exercise in structural induction over Definition 1.1 to see that it has the following general form, for some $1 \leq i \leq \mathcal{N}(G)$:

$$
\mathcal{R}_\alpha(G,\, \mathbf{v},\, H,\, Q) = r_\alpha(Q) v_i \times \prod_{n=1}^{\mathcal{D}_\alpha(G)} \frac{\min(a_n(\mathbf{v}),\, b_n(\mathbf{v}))}{a_n(\mathbf{v})} \tag{C.1}
$$

where for any $1 \leq n \leq \mathcal{D}_\alpha(G)$:

$$
a_n(\mathbf{v}) \geq r_\alpha(Q) v_i \times \prod_{m=1}^{n-1} \frac{\min(a_m(\mathbf{v}),\, b_m(\mathbf{v}))}{a_m(\mathbf{v})}
$$

Now, the functions $a_n(\cdot)$ and $b_n(\cdot)$ are just instances of apparent rate (Definition 1.2). So they and their minimum, $\min(a_n(\cdot), b_n(\cdot))$, are all piecewise-linear on closed subsets of $\mathbb{R}_+^{\mathcal{N}(G)}$, each defined by a system of linear inequalities. These subsets thus form a covering of $\mathbb{R}_+^{\mathcal{N}(G)}$ by closed convex sets. Take an arbitrary such region, say $A \subseteq \mathbb{R}_+^{\mathcal{N}(G)}$. For $\mathbf{v} \in A$, some of the terms of the product in Equation (C.1) will cancel and, re-ordering indices where necessary, for some $D \leq \mathcal{D}_\alpha(G)$:

$$\mathcal{R}_\alpha(G,\, \mathbf{v},\, H,\, Q) = r_\alpha(Q)v_i \times \prod_{n=1}^{D} \frac{b_n(\mathbf{v})}{a_n(\mathbf{v})}$$

and for any $1 \leq n \leq D$, the following two inequalities hold:

$$\begin{aligned} a_n(\mathbf{v}) &\geq r_\alpha(P)v_i \times \prod_{m=1}^{n-1} \frac{b_m(\mathbf{v})}{a_m(\mathbf{v})} \\ a_n(\mathbf{v}) &\geq b_n(\mathbf{v}) \end{aligned} \tag{C.2}$$

Furthermore, on the interior of A, $\text{int}(A)$, $\mathcal{R}_\alpha(G,\, \cdot,\, H,\, Q)$ is differentiable since the $a_n(\cdot)$ and $b_n(\cdot)$ are linear. So for any $1 \leq j \leq \mathcal{N}(G)$ and $\mathbf{v} \in \text{int}(A)$:

$$\frac{\partial \mathcal{R}_\alpha(G,\, \cdot,\, H,\, Q)}{\partial v_j}(\mathbf{v}) = \begin{cases} r_\alpha(Q) \times \prod_{n=1}^{D} \frac{b_n(\mathbf{v})}{a_n(\mathbf{v})} + r_\alpha(Q)v_i \times \frac{\partial}{\partial v_j}\left[\prod_{n=1}^{D} \frac{b_n(\mathbf{v})}{a_n(\mathbf{v})}\right] & : j = i \\ r_\alpha(Q)v_i \times \frac{\partial}{\partial v_j}\left[\prod_{n=1}^{D} \frac{b_n(\mathbf{v})}{a_n(\mathbf{v})}\right] & : j \neq i \end{cases}$$

Write $F[l](\mathbf{v}) := \prod_{n=1}^{l} \frac{b_n(\mathbf{v})}{a_n(\mathbf{v})}$ for $0 \leq l \leq D$. Then:

$$r_\alpha(Q)v_i \times \frac{\partial F[l]}{\partial v_j}(\mathbf{v}) = r_\alpha(Q)v_i \frac{F[l-1](\mathbf{v})}{a_l(\mathbf{v})} \frac{\partial b_l}{\partial v_j}(\mathbf{v}) - r_\alpha(Q)v_i \frac{F[l](\mathbf{v})}{a_l(\mathbf{v})} \frac{\partial a_l}{\partial v_j}(\mathbf{v}) + \frac{b_l(\mathbf{v})}{a_l(\mathbf{v})} r_\alpha(Q)v_i \times \frac{\partial F[l-1]}{\partial v_j}(\mathbf{v})$$

Applying the inequalities of Equation (C.2), we obtain:

$$\left| r_\alpha(Q)v_i \times \frac{\partial F[l]}{\partial v_j}(\mathbf{v}) \right| \leq \left| \frac{\partial b_l}{\partial v_j}(\mathbf{v}) \right| + \left| \frac{\partial a_l}{\partial v_j}(\mathbf{v}) \right| + \left| r_\alpha(Q)v_i \times \frac{\partial F[l-1]}{\partial v_j}(\mathbf{v}) \right|$$

It is clear from the definition of apparent rate (Section A.1) that only one of $\left|\frac{\partial b_l}{\partial v_j}(\mathbf{v})\right|$ and $\left|\frac{\partial a_l}{\partial v_j}(\mathbf{v})\right|$ can be non-zero, and it is no greater than $r_\alpha(Q')$ for some standard PEPA component, Q'. Thus by induction:

$$\sup_{\mathbf{v} \in \text{int}(A)} \left| \frac{\partial \mathcal{R}_\alpha(G,\, \cdot,\, H,\, Q)}{\partial v_j}(\mathbf{v}) \right| \leq (\mathcal{D}_\alpha(G) + 1)\mathcal{Q}_\alpha^{\max}(G)$$

Now considering all action types in the same way, we have:

$$\sup_{\mathbf{v} \in \text{int}(A)} \left| \frac{\partial f_k}{\partial v_j}(\mathbf{v}) \right| \leq 2 \sum_{\alpha_i \in \mathcal{A}} (\mathcal{D}_{\alpha_i}(G) + 1)\mathcal{Q}_{\alpha_i}^{\max}(G)$$

We now apply Lemma C.1 on the open convex set, $\text{int}(A)$ to show that $\mathbf{f}(\cdot)$ is Lipschitz continuous on A with a Lipschitz constant, $\mathcal{K}(G) := 2\mathcal{N}(G) \sum_{\alpha_i \in \mathcal{A}} (\mathcal{D}_{\alpha_i}(G) + 1)\mathcal{Q}_{\alpha_i}^{\max}(G)$.

For general $\mathbf{v}_1,\, \mathbf{v}_2 \in \mathbb{R}_+^{\mathcal{N}(G)}$, consider the line connecting them, $(1-t)\mathbf{v}_1 + t\mathbf{v}_2$ for $t \in [0,\, 1]$. Let the closed convex sets making up the covering of $\mathbb{R}_+^{\mathcal{N}(G)}$ defined above be $\{A_i\}_{i=1}^{N}$. Assume the line between $\mathbf{v}_1$ and $\mathbf{v}_2$ intersects $k \geq 1$ of them, then re-ordering where necessary, there exist $\{t_j\}_{j=1}^{k+1}$ with $t_1 = 0$, $t_{k+1} = 1$ such

that for each $1 \leq j < k$, $t_j < t_{j+1}$ and $(1-t)\mathbf{v}_1 + t\mathbf{v}_2 \in A_j$ for $t \in [t_j, t_{j+1}]$. Now write:

$$\begin{aligned}
\|\mathbf{f}(\mathbf{v}_1) - \mathbf{f}(\mathbf{v}_2)\| &= \left\| \sum_{j=1}^{k} (\mathbf{f}((1-t_j)\mathbf{v}_1 + t_j\mathbf{v}_2) - \mathbf{f}((1-t_{j+1})\mathbf{v}_1 + t_{j+1}\mathbf{v}_2)) \right\| \\
&\leq \sum_{j=1}^{k} \|(\mathbf{f}((1-t_j)\mathbf{v}_1 + t_j\mathbf{v}_2) - \mathbf{f}((1-t_{j+1})\mathbf{v}_1 + t_{j+1}\mathbf{v}_2))\| \\
&\leq \mathcal{K}(G) \sum_{j=1}^{k} \|((1-t_j)\mathbf{v}_1 + t_j\mathbf{v}_2) - ((1-t_{j+1})\mathbf{v}_1 + t_{j+1}\mathbf{v}_2)\| \\
&= \mathcal{K}(G)\|\mathbf{v}_1 - \mathbf{v}_2\|
\end{aligned}$$

as required. □

C.2. Proof of Theorem 3.1

Proof. We require to show for all $t, s \geq 0$, that $\mathbb{E}[\bar{\mathbf{M}}(s+t) \,|\, \mathcal{F}_s] = \bar{\mathbf{M}}(s)$ a.s., where $\mathcal{F}_s$ is the natural filtration of $\bar{\mathbf{N}}(s)$. Since $\mathbb{E}[\bar{\mathbf{M}}(s+t) \,|\, \mathcal{F}_s] = \bar{\mathbf{M}}(s) + \mathbb{E}[\bar{\mathbf{M}}(t)]$ using the Markov property, this is equivalent to showing that $\mathbb{E}[\bar{\mathbf{M}}(t)] = 0$ for all $t \geq 0$.

Let the jump times of $\bar{\mathbf{N}}(t)$ be $\{\tau_j\}_{j=0}^{\infty}$ with $\tau_0 := 0$. Now consider $\bar{\mathbf{M}}(t)$ stopped at τ_1, i.e. $\bar{\mathbf{M}}(t \wedge \tau_1)$. Then using homogeneity of $\mathbf{f}(\cdot)$ (Lemma 1.6), that is, the fact that $\mathbf{f}(\mathbf{N}(t)) = \mathcal{S}(G)\mathbf{f}(\bar{\mathbf{N}}(t))$, we have:

$$\mathbb{E}[\bar{\mathbf{M}}(t \wedge \tau_1)] = -\mathbb{E}[1_{\{0 \leq t < \tau_1\}} \cdot t\mathbf{f}(\bar{\mathbf{N}}(0))] + \mathbb{E}[1_{\{t \geq \tau_1\}} \cdot (\bar{\mathbf{N}}(\tau_1) - \tau_1 \mathbf{f}(\bar{\mathbf{N}}(0)))] = 0$$

by a straightforward argument. Repeating the argument using the Markov property, we see that $\mathbb{E}[\bar{\mathbf{M}}(t \wedge \tau_j)] = 0$ for all $j \geq 0$. Therefore, we may write for all $j \geq 0$, $\mathbb{E}[\bar{\mathbf{M}}(t)] = \mathbb{E}[1_{\{t \geq \tau_j\}} \cdot (\bar{\mathbf{M}}(t) - \bar{\mathbf{M}}(\tau_j))]$, and also in the limit:

$$\mathbb{E}[\bar{\mathbf{M}}(t)] = \lim_{j \to \infty} \mathbb{E}[1_{\{t \geq \tau_j\}} \cdot (\bar{\mathbf{M}}(t) - \bar{\mathbf{M}}(\tau_j))] = \lim_{j \to \infty} \mathbb{E}[1_{\{t \geq \tau_j\}} \cdot \bar{\mathbf{M}}(t)] - \lim_{j \to \infty} \mathbb{E}[1_{\{t \geq \tau_j\}} \cdot \bar{\mathbf{M}}(\tau_j)] \tag{C.3}$$

We can bound τ_j in distribution by an Erlang random variable with parameters k and R, where R is the finite maximum jump rate of $\bar{\mathbf{N}}(t)$. So $\lim_{j \to \infty} \mathbb{E}[1_{\{t \geq \tau_j\}}] = 0$. Then the first term of Equation (C.3) is zero by monotone convergence and the second is zero by the Cauchy-Schwarz inequality. □

C.3. Proof of Theorem 3.2

Proof. Recall that we are considering the grouped PEPA model, G, with associated rescaled aggregated stochastic process, $\bar{\mathbf{N}}(t)$ and system of ODEs, $\dot{\bar{\mathbf{v}}}(t) = \mathbf{f}(\bar{\mathbf{v}}(t))$.

Recall from Lemma 1.5 that a Lipschitz constant of the function $\mathbf{f}(\cdot)$ is $\mathcal{K}(G)$. Now note that by definition and the triangle inequality, we have for all $t \in [0, \infty)$:

$$\begin{aligned}
\|\bar{\mathbf{N}}(t) - \bar{\mathbf{v}}(t)\| &= \|\bar{\mathbf{M}}(t) - \bar{\mathbf{v}}(t) + \int_0^t \mathbf{f}(\bar{\mathbf{N}}(s))\, ds\| \\
&\leq \|\bar{\mathbf{M}}(t)\| + \int_0^t \|\mathbf{f}(\bar{\mathbf{N}}(s)) - \mathbf{f}(\bar{\mathbf{v}}(s)\|\, ds \\
&\leq \|\bar{\mathbf{M}}(t)\| + \mathcal{K}(G) \int_0^t \|\bar{\mathbf{N}}(s) - \bar{\mathbf{v}}(s)\|\, ds
\end{aligned}$$

For $\delta = \epsilon \exp(-\mathcal{K}(G)T)$, define the event $A := \left\{ \sup_{t \in [0,\, T]} \|\bar{\mathbf{M}}(t)\| > \delta \right\}$. On the complement of this event, we may apply Grönwall's inequality [23, Pg. 498] to the function, $e(s) := \sup_{t \in [0,\, s]} \|\bar{\mathbf{N}}(s) - \bar{\mathbf{v}}(s)\|$, to obtain,

$\sup_{t\in[0,\,T]} \|\bar{\mathbf{N}}(t) - \bar{\mathbf{v}}(t)\| \leq \epsilon$. Thus we have shown the following result:

$$\mathbb{P}\left\{\sup_{t\in[0,\,T]} \|\bar{\mathbf{N}}(t) - \bar{\mathbf{v}}(t)\| > \epsilon\right\} \leq \mathbb{P}(A)$$

So it remains to bound $\mathbb{P}(A)$. As in the proof of Theorem 3.1, let the jump times of $\bar{\mathbf{N}}(t)$ be $\{\tau_j\}_{j=0}^{\infty}$ and note that for any $k \geq 0$:

$$A = \left\{\sup_{t\in[0,\,T]} \|\bar{\mathbf{M}}(t)\|^2 > \delta^2\right\} = \{\tau_k < T\} \cup \left\{\sup_{t\in[0,\,T]} \|\bar{\mathbf{M}}(t\wedge\tau_k)\|^2 > \delta^2\right\}$$

where $\bar{\mathbf{M}}(t\wedge\tau_k)$ is $\bar{\mathbf{M}}(t)$ stopped at τ_k, and is also a martingale by the optional stopping theorem. So we can apply Doob's L^2-martingale inequality to it to obtain:

$$\mathbb{P}\left\{\sup_{t\in[0,\,T]} \|\bar{\mathbf{M}}(t\wedge\tau_k)\|^2 > \delta^2\right\} \leq \delta^{-2}\mathbb{E}\left[\sup_{t\in[0,\,T]} \|\bar{\mathbf{M}}(t\wedge\tau_k)\|^2\right] \leq 4\delta^{-2}\mathbb{E}\left[\|\bar{\mathbf{M}}(T\wedge\tau_k)\|^2\right]$$

Thus we have, for any $k \geq 0$, $\mathbb{P}(A) \leq \mathbb{P}\{\tau_k < T\} + 4\delta^{-2}\mathbb{E}\left[\|\bar{\mathbf{M}}(T\wedge\tau_k)\|^2\right]$.

Choose some $q > \mathcal{Q}^{\max}(G)$, such that qT is an integer. Then we can fix integer $k = \mathcal{S}(G)qT$. Now τ_k is bounded below in distribution by the sum of k independent and identically exponentially-distributed random variables each with parameter the maximal jump rate of $\bar{\mathbf{N}}(t)$, $\mathcal{S}(G)\mathcal{Q}^{\max}(G)$. Denote this Erlang random variable by μ_k. So we have:

$$\mathbb{P}\{\tau_k < T\} \leq \mathbb{P}\{\mu_k < T\} = \frac{\gamma(\mathcal{S}(G)qT,\, \mathcal{S}(G)\mathcal{Q}^{\max}(G)T)}{(\mathcal{S}(G)qT-1)!}$$

where $\gamma(\cdot,\,\cdot)$ is the lower incomplete gamma function.

Now to bound $\mathbb{E}\left[\|\bar{\mathbf{M}}(T\wedge\tau_k)\|^2\right]$, note that:

$$\begin{aligned}\mathbb{E}\left[\|\bar{\mathbf{M}}(T\wedge\tau_k)\|^2\right] &= \mathbb{E}[\mathbf{1}_{\{T\geq\tau_k\}}\|\bar{\mathbf{M}}(\tau_k)\|^2] + \sum_{i=0}^{k-1}\mathbb{E}[\mathbf{1}_{\{\tau_i\leq T<\tau_{i+1}\}}\|\bar{\mathbf{M}}(\tau_i)\|^2] \\ &\leq \mathbb{P}\{T\geq\tau_k\}\mathbb{E}[\|\bar{\mathbf{M}}(\tau_k)\|^2] + \sum_{i=0}^{k-1}\mathbb{P}\{\tau_i\leq T<\tau_{i+1}\}\mathbb{E}[\|\bar{\mathbf{M}}(\tau_i)\|^2]\end{aligned} \tag{C.4}$$

For any $i \geq 0$, $\mathbb{E}[\|\bar{\mathbf{M}}(\tau_i)\|^2] \leq \sum_{j=0}^{i-1}\mathbb{E}[\|\bar{\mathbf{M}}(\tau_{j+1}) - \bar{\mathbf{M}}(\tau_j)\|^2]$, and for any $j \geq 0$:

$$\begin{aligned}&\mathbb{E}[\|\bar{\mathbf{M}}(\tau_{j+1}) - \bar{\mathbf{M}}(\tau_j)\|^2] \\ &= \sum_{n=1}^{\mathcal{N}(G)}\mathbb{E}[(\bar{N}_n(\tau_{j+1}) - \bar{N}_n(\tau_j))^2 + (\tau_{j+1}-\tau_j)^2 f_n^2(\bar{\mathbf{N}}(\tau_j)) - 2(\bar{N}_n(\tau_{j+1}) - \bar{N}_n(\tau_j))(\tau_{j+1}-\tau_j)f_n(\bar{\mathbf{N}}(\tau_j))]\end{aligned} \tag{C.5}$$

For any possible state of the CTMC, $\bar{\mathbf{N}}(t)$, say $\bar{N}$, write $r(\bar{N})$ for the sum of all of the rates of all outgoing transitions from that state. Then, for any possible state, $\bar{N}$, we have, by homogeneity of $r(\cdot)$ (Lemma 1.6):

$$\begin{aligned}\mathbb{E}[2(\bar{N}_n(\tau_{j+1}) - \bar{N}_n(\tau_j))(\tau_{j+1}-\tau_j)f_n(\bar{\mathbf{N}}(\tau_j)) \,|\, \bar{\mathbf{N}}(\tau_j) = \bar{N}] &= 2f_n^2(\bar{N})/(\mathcal{S}^2(G)r^2(\bar{N})) \\ &= \mathbb{E}[(\tau_{j+1}-\tau_j)^2 f_n^2(\bar{\mathbf{N}}(\tau_j)) \,|\, \bar{\mathbf{N}}(\tau_j) = \bar{N}]\end{aligned}$$

Thus, combining Equations (C.4) and (C.5), we obtain:

$$\mathbb{E}\left[\|\bar{\mathbf{M}}(T\wedge\tau_k)\|^2\right] \leq k\sum_{n=1}^{\mathcal{N}(G)}\mathbb{E}[(\bar{N}_n(\tau_{k+1}) - \bar{N}_n(\tau_k))^2] \leq k\mathcal{N}(G)/\mathcal{S}^2(G) = qT\mathcal{N}(G)/\mathcal{S}(G)$$

since each component type count changes by at most one component at each jump time. The required bound follows.

To show that the bound tends to zero as $\mathcal{S}(G) \to \infty$ and everything else remains fixed, it remains just to show that:

$$\mathbb{P}\{\mu_k < T\} = \frac{\gamma(\mathcal{S}(G)qT,\, \mathcal{S}(G)\mathcal{Q}^{\max}(G)T)}{(\mathcal{S}(G)qT-1)!} \to 0$$

as $\mathcal{S}(G) \to \infty$ with fixed q, T and $\mathcal{Q}^{\max}(G)$. Now for any $a > T$:

$$\mathbb{P}\{\mu_k < T\} \leq \mathbb{P}\{(\mu_k - a)^2 \geq (a-T)^2\} = \mathbb{P}\{\mu_k^2 - 2a\mu_k \geq T^2 - 2aT\} \leq \frac{1}{T^2-2aT}(\mathbb{E}[\mu_k^2] - 2a\mathbb{E}[\mu_k])$$

Now choose $a = \mathbb{E}[\mu_k] = \frac{q}{Q}T > T$ and note that $\mathrm{Var}[\mu_k] = \frac{qT}{\mathcal{S}(G)Q^2}$, so $\mathbb{P}\{\mu_k < T\} \leq \frac{1}{T(T-q/Q)}\frac{qT}{\mathcal{S}(G)Q^2}$ and the required result follows. □

C.4. Proof of Theorem 3.4

Proof. Fix $T > \gamma$ and $\epsilon < \min(\gamma, T-\gamma)$. Choose $\delta^- := \mathbf{p}.\bar{\mathbf{v}}(\gamma-\epsilon) - P > 0$ and $\delta^+ := P - \mathbf{p}.\bar{\mathbf{v}}(\gamma+\epsilon) > 0$. Then:

$$\begin{aligned}
\mathbb{P}\{|\sigma_i - \gamma| > \epsilon\} &= \mathbb{P}(\{\sigma_i - \gamma > \epsilon\} \cup \{\gamma - \sigma_i > \epsilon\}) \\
&\leq \mathbb{P}(\left\{\|\bar{\mathbf{N}}^{G_i}(\gamma-\epsilon) - \bar{\mathbf{v}}(\gamma-\epsilon)\| \geq \delta^-\right\} \cup \{\|\bar{\mathbf{N}}^{G_i}(\gamma+\epsilon) - \bar{\mathbf{v}}(\gamma+\epsilon)\| \geq \delta^+\}) \\
&\leq \mathbb{P}\left\{\sup_{t\in[0,\,T]} \|\bar{\mathbf{N}}^{G_i}(t) - \bar{\mathbf{v}}(t)\| \geq \min(\delta^-,\, \delta^+)\right\}
\end{aligned}$$

Now by structural equivalence, the quantities, q, $\mathcal{Q}^{\max}(G)$, $|\mathcal{B}|$ and $\mathcal{K}(G)$ in Theorem 3.2 defined for each grouped PEPA model, G_i, are independent of i. So by that theorem, if $\mathcal{S}(G_i) \to \infty$ as $i \to \infty$, we have the desired result. □

C.5. Proof of Theorem 3.5

Proof. Now, for any $k > 0$, $\delta > 0$ and $t \in [0, T]$:

$$\begin{aligned}
\frac{1}{k^2\mathcal{S}^2(G_i)}\|\mathbb{E}[\mathbf{N}^{G_i}(t)] - \mathbf{v}^{G_i}(t)\|^2 = \frac{1}{k^2}\|\mathbb{E}[\bar{\mathbf{N}}^{G_i}(t)] - \bar{\mathbf{v}}(t)\|^2 &\leq \frac{1}{k^2}\mathbb{E}[\|\bar{\mathbf{N}}^{G_i}(t) - \bar{\mathbf{v}}(t)\|^2] \\
&\leq \frac{1}{k^2}\left[\mathbb{P}\left\{\sup_{t\in[0,\,T]} \|\bar{\mathbf{N}}^{G_i}(t) - \bar{\mathbf{v}}(t)\|^2 \geq \delta^2\right\} 4|\mathcal{B}| + \delta^2\right]
\end{aligned}$$

By Theorem 3.2, the limit of this quantity as $i \to \infty$ is δ^2/k^2. Since this holds for any $\delta > 0$, we have, $1/(k\mathcal{S}(G_i))\|\mathbb{E}[\mathbf{N}^{G_i}(t)] - \mathbf{v}^{G_i}(t)\| \to 0$ as $i \to \infty$. This gives the first result. The second follows too since $1/(k\mathcal{S}(G_i)+1)\|\mathbb{E}[\mathbf{N}^{G_i}(t)] - \mathbf{v}^{G_i}(t)\| < 1/(k\mathcal{S}(G_i))\|\mathbb{E}[\mathbf{N}^{G_i}(t)] - \mathbf{v}^{G_i}(t)\|$. □

C.6. Lemmas

Lemma C.1. *Let $A \subseteq \mathbb{R}^n$ be convex and open. Let $\mathbf{g} : \bar{A} \to \mathbb{R}^n$ be a function continuous on $\bar{A}$ and differentiable on A. Assume also that for $i, j \in \{1, \ldots, n\}$:*

$$\sup_{\mathbf{x}\in A}\left|\frac{\partial g_i}{\partial x_j}(\mathbf{x})\right| \leq \Lambda < \infty$$

Then $\mathbf{g}$ is Lipschitz continuous on $\bar{A}$ with a Lipschitz constant, $n\Lambda$.

Proof. Let $\mathbf{x}$, $\mathbf{y} \in A$ be arbitrary, and define the function $\mathbf{G} : [0,\, 1] \to \mathbb{R}^n$ by $\mathbf{G}(t) := \mathbf{g}((1-t)\mathbf{x} + t\mathbf{y})$. Now, by convexity of A, $\mathbf{G}$ is differentiable on $(0,\, 1)$ and we have for all $t \in (0,\, 1)$, $\mathbf{G}'(t) = \mathbf{D}\mathbf{g}((1-t)\mathbf{x}+t\mathbf{y})\cdot(\mathbf{y}-\mathbf{x})$. Then:

$$|G_i(1) - G_i(0)|^2 = \left[\int_0^1 \frac{dG_i}{dt}(s)\, ds\right]^2 = \left[\sum_{j=1}^n (y_j - x_j)\right]^2 \left[\int_0^1 \frac{\partial g_i}{\partial x_j}((1-s)\mathbf{x} + s\mathbf{y})\, ds\right]^2$$
$$\leq n\|\mathbf{y} - \mathbf{x}\|^2 \Lambda^2$$

by the Cauchy–Schwarz inequality. So:

$$\|\mathbf{g}(\mathbf{y}) - \mathbf{g}(\mathbf{x})\| = \|\mathbf{G}(1) - \mathbf{G}(0)\| \leq n\Lambda\|\mathbf{y} - \mathbf{x}\|$$

as required. The extension to $\bar{A}$ is trivial by continuity of $\mathbf{g}$ and continuity of norms. □