Theoretical Computer Science 413 (2012) 106–141

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Fluid computation of passage-time distributions in large Markov models

Richard A. Hayden, Anton Stefanek, Jeremy T. Bradley *

Department of Computing, Imperial College London, Huxley Building, 180 Queen's Gate, London SW7 2BZ, United Kingdom

ARTICLE INFO

Keywords:
Fluid approximation
Passage-time distribution
Stochastic process algebra
Service level agreement
Scalability analysis

ABSTRACT

Recent developments in the analysis of large Markov models facilitate the fast approximation of transient characteristics of the underlying stochastic process. *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, global 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 upper and lower approximations for the entire cumulative distribution function of these passage-time random variables, using moment-based techniques. Additionally, we show that, for passage-time measures involving individual components, the cumulative distribution function can be directly approximated by fluid techniques.

Finally, using the GPA tool, we take advantage of the rapid fluid computation of passage times to show how a multi-class client–server system can be optimised to satisfy multiple service level agreements.

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 virtualised web service should process a request within 0.6 s, 98% 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 here massively-parallel Markov models that consist of synchronising groups of component Markov chains, phrased in the *grouped PEPA (GPEPA)* process algebra [2]. Fluid analysis involves approximating their underlying discrete state space with continuous real-valued variables and describing the transient evolution of those variables with ordinary differential equations (ODEs). The solution to the ODEs is an approximation to discrete stochastic processes which count the number of Markov chain components in the model which are in a given state. The fluid analysis framework for GPEPA [2] built on the original approach for PEPA by Hillston [3]. Similar approaches have also been developed

* Corresponding author. Tel.: +44 2075941234.
E-mail addresses: rh@doc.ic.ac.uk (R.A. Hayden), as1005@doc.ic.ac.uk (A. Stefanek), jb@doc.ic.ac.uk (J.T. Bradley).
URLs: http://www.doc.ic.ac.uk/~rh (R.A. Hayden), http://www.doc.ic.ac.uk/~as1005 (A. Stefanek), http://www.doc.ic.ac.uk/~jb (J.T. Bradley).

0304-3975/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2011.07.017

for various other stochastic process algebras such as *stochastic concurrent constraint programming (sCCP)* [4] and *stochastic π-calculus* [5], and *stochastic Petri net* formalisms [6]. These approaches are also very similar to the so-called *mean-field* methods of Le Boudec et al. [7], Bobbio et al. [8] and Bakhshi et al. [9].

The goal of this paper is to show that fluid-analysis techniques can be used to compute passage-time distributions efficiently. In terms of previous work which has also considered passage times in the context of fluid analysis, Tribastone [10] has shown how *average* steady-state individual passage times in PEPA models may be computed by combining fluid analysis with the classical result from queueing theory known as Little's Law [11]. Such an approach has also been applied by Ding [12] and is used by Clark et al. [13] and Kesidis et al. [14].

Our approach however builds on the preliminary work of Bradley et al. [15], which noted that a certain class of conditional passage-time measure was equivalent to the *time to extinction* of a certain set of components within a modified model. 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. As we will see, this perspective allows us to develop techniques which give access to the *distribution* of passage-time measurements rather than just averages. Such an ability is key since most SLAs are specified in terms of passage-time quantiles rather than averages. Our first contribution is to develop the ideas of [15] to introduce two new classes of passage-time measure which are amenable to this form of approximation: *global* and *individual* passage times.

In Section 3.1, we introduce 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. Specifically, in Section 3.1.1, we present a deterministic point-mass approximation, to which an appropriate sequence of global passage times will converge as the component populations increase. Where the component populations are not large enough for the deterministic approximation to be accurate, we improve upon this significantly by showing how efficient approximate upper and lower bounds on the cumulative distribution function of the entire passage time can be derived (Section 3.1.2).

Section 3.2 shows how *individual passage times* can track the evolution of single components in massively parallel systems, both in the steady-state and transient regimes. For these individual passage times we show how the entire cumulative distribution function can be well approximated by fluid techniques. Individual passage-time measures are analogous to *tagged-customer* measures in stochastic Petri nets [16,17]. However these approaches still rely on traditional explicit state Markov chain analysis and are thus susceptible to state-space explosion.

In Section 4, we provide convergence proofs for both global and individual passage times. The convergence proof for so-called *steady-state individual passage times* requires a proof of convergence of the fluid approximation in the steady-state regime (Theorem 4.3) which is a powerful new generic result in itself extending similar discrete-time results of Benaïm and Le Boudec [7] to the continuous-time case. We also provide for the first time inexpensive methods for the verification of the asymptotic stability of the approximating differential equations' fixed point which is a precondition of Theorem 4.3 (Sections 4.2.1 and 4.2.2). Finally, we demonstrate an implementation of the passage time techniques given in this paper using the GPA tool [18] (can be downloaded from http://code.google.com/p/gpanalyser). Section 5 contains a worked example of a multi-class client–server model. This model exploits the rapid fluid computation of passage times possible in GPA to show how SLAs for each class of customer can be individually satisfied while minimising the number of servers required in the system.

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

1.1. Grouped PEPA

Grouped PEPA [2] (or *GPEPA*) is a simple extension of the stochastic process algebra PEPA [19], which facilitates the application of fluid analysis techniques to massively-parallel models. A GPEPA model consists of a number of labelled cooperating component groups, each of which consists of a large number of components operating together in parallel. We refer to the components within these groups as *fluid components*. The fluid components are those whose state will be tracked explicitly by an approximating system of differential equations.

1.1.1. Component groups

A *component group* is a parallel cooperation of a normally large number of fluid components. By *parallel* here we mean that there is no synchronisation between individual members of the component group. Syntactically, a component group, D, is specified by the following grammar:

$$D ::= D \between D \mid P \tag{1.1}$$

where P is a *fluid component*, a PEPA process algebra term to be formally introduced in the next section. The combinator $\between$ represents parallel, unsynchronised cooperation between fluid components.

As we will see in more detail in Section 1.1.3, a *grouped PEPA model* is formed by combining multiple labelled component groups together. Syntactically, the grammar for a grouped PEPA model G is:

$$G ::= G \underset{L}{\bowtie} G \mid Y\{D\} \tag{1.2}$$

Prefix

$$\frac{}{(\alpha, r).P \xrightarrow{(\alpha, r)} P}$$

Competitive Choice

$$\frac{P \xrightarrow{(\alpha, r)} P'}{P + Q \xrightarrow{(\alpha, r)} P'} \qquad \frac{Q \xrightarrow{(\alpha, r)} Q'}{P + Q \xrightarrow{(\alpha, r)} Q'}$$

Cooperation

$$\frac{P \xrightarrow{(\alpha, r)} P'}{P \underset{S}{\bowtie} Q \xrightarrow{(\alpha, r)} P' \underset{S}{\bowtie} Q} \quad (\alpha \notin S) \qquad \frac{Q \xrightarrow{(\alpha, r)} Q'}{P \underset{S}{\bowtie} Q \xrightarrow{(\alpha, r)} P \underset{S}{\bowtie} Q'} \quad (\alpha \notin S)$$

$$\frac{P \xrightarrow{(\alpha, r_1)} P' \quad Q \xrightarrow{(\alpha, r_2)} Q'}{P \underset{S}{\bowtie} Q \xrightarrow{(\alpha, R)} P' \underset{S}{\bowtie} Q'} \quad (\alpha \in S)$$

where $R := \frac{r_1}{r_\alpha(P)} \frac{r_2}{r_\alpha(Q)} \min(r_\alpha(P), r_\alpha(Q))$

Hiding

$$\frac{P \xrightarrow{(\alpha, r)} P'}{P/H \xrightarrow{(\alpha, r)} P'/H} \quad (\alpha \notin H) \qquad \frac{P \xrightarrow{(\alpha, r)} P'}{P/H \xrightarrow{(\tau, r)} P'/H} \quad (\alpha \in H)$$

Constant

$$\frac{P \xrightarrow{(\alpha, r)} P'}{\boldsymbol{C} \xrightarrow{(\alpha, r)} P'} \quad (\boldsymbol{C} \stackrel{def}{=} P)$$

Fig. 1. Structured operational semantics of PEPA.

where Y is a group label, unique to each component group. The term $G \underset{L}{\bowtie} G$ represents synchronisation over the set of action types $L \subseteq \mathcal{A} \setminus \{\tau\}$. Exactly what this means in terms of the operational semantics of PEPA will be discussed in the following two sections. Informally, a fluid component in one component group may synchronise with a fluid component in *another* component group, but, as mentioned above, fluid components may not synchronise with other fluid components in the same component group. We will see that this restriction defines a class of models to which fluid analysis is naturally applicable.

1.1.2. Fluid components in GPEPA

A fluid component is simply defined to be any standard PEPA process algebra component. Syntactically, a fluid component is specified by the standard PEPA grammar [19]:

$$\begin{aligned} S &::= (\alpha, r).S \mid S + S \mid C_S \\ P &::= P \underset{L}{\bowtie} P \mid P/L \mid S \mid C_P \end{aligned} \tag{1.3}$$

where $\alpha \in \mathcal{A}$ is an *action type*, $L \subseteq \mathcal{A} \setminus \{\tau\}$ and $r \in \mathbb{R}_+ \cup \{n\top \mid n \in \mathbb{Q}, n > 0\}$ is a rate parameter. A timed transition is also referred to as an activity.

In line with Eq. (1.3), a fluid component can be a purely *sequential component*, S, or a *model component*, P, with its own internal parallelism. C_S and C_P represent constants which denote sequential components or model components respectively. The effect of this syntactic separation between constants is to constrain legal PEPA components to be only cooperations between sequential components.

Now we introduce informally the intended semantics of the PEPA syntax defined above. The formal structured operational semantics are given in Fig. 1.

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. $(\alpha, r).P$ 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. If an activity in P completes first, the system then proceeds by taking on the behaviour of the derivative of P following the completed action; and vice-versa for Q.

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, written as $\boldsymbol{X} \stackrel{def}{=} P$.

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 action types 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.

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 action types 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. Formally, for a given action type $\alpha \in \mathcal{A}$, it is thus calculated by summing the rates of all enabled activities of this type $r_\alpha(P) := \sum_{P \xrightarrow{(\alpha,\lambda)}} \lambda$. Apparent rate can also be defined equivalently in a recursive manner over the PEPA grammar, see e.g. [19].

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.

In some cases, when the rate of a shared activity is determined by only one component in the cooperation, then the other component is defined as *passive* with respect to that activity. This means that the rate of the activity is left unspecified (denoted $\top$) and is determined upon cooperation, by the rate of the activity in the other component. In defining fluid components, we restrict all passive actions to be synchronised in a final (outermost) fluid component, so as not to allow passive cooperation between component groups. Also a fluid component is not allowed to offer the same action type both passively and actively (a standard restriction in PEPA).

For a given fluid component P, its *derivative set* $ds(P)$ is the set of components reachable from P. That is, $ds(P)$ is the smallest set of components such that $P \in ds(P)$ and if for any $P_1 \in ds(P)$, $P_1 \xrightarrow{(\alpha,r)} P_2$ then $P_2 \in ds(P)$. We also define the *multiset* of enabled activities of a PEPA component P to be $\mathcal{A}ct(P) := \{\!|(\alpha, r) : P \xrightarrow{(\alpha,r)} |\!\}$, and write $\mathcal{A}(P)$ for the set of action types which are used by any derivative state of P, that is, $\mathcal{A}(P) := \cup_{P' \in ds(P)}\{\alpha : P' \xrightarrow{(\alpha,\cdot)}\}$. The *derivation graph* of P is a labelled and directed multigraph whose nodes are the derivative states of P and two nodes in the multigraph, say P_1 and $P_2 \in ds(P)$, have a directed arc between them for every transition $P_1 \xrightarrow{(\alpha,\lambda)} P_2$. The derivation graph can then be interpreted naturally as a CTMC, whose states are given by the derivative states and each arc represents a transition at the rate of the activity labelling the arc. We call this the *underlying CTMC* of P.

1.1.3. Grouped PEPA examples

To illustrate more clearly how component groups and fluid components are used together to construct grouped PEPA models, we will now introduce a PEPA model which will also serve as a running example throughout this paper. The type of model we wish to consider is one which exhibits massive parallelism. We present such a system below where 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 fetch the data for the client. At any stage, a server in the pool may fail. Clients may also timeout when waiting for data after their initial request. Classical Markov chain analysis of any variety requires exploration of the global state space and, even for such a simple system, we will see that this quickly becomes computationally infeasible.

We capture this scenario of n clients cooperating on the *request* and *data* actions with m resources with the following PEPA system equation:

$$\boldsymbol{Client}[n] \underset{L}{\bowtie} \boldsymbol{Server}[m]$$

where $L = \{request, data\}$ and $C[n]$ represents n parallel copies of component C:

$$C[n] := \underbrace{(C \parallel \ldots \parallel C)}_{n} \tag{1.4}$$

Each client is represented as a ***Client*** component and each server as a ***Server*** component. Each client operates forever in a loop, completing three tasks in sequence: *request*, *data* and then *think*; and they may also perform a *timeout* action when waiting for data:

$$\textbf{\textit{Client}} \stackrel{def}{=} (request, r_r).\textbf{\textit{Client\_waiting}}$$

$$\textbf{\textit{Client\_waiting}} \stackrel{def}{=} (data, r_d).\textbf{\textit{Client\_think}} + (timeout, r_{tmt}).\textbf{\textit{Client}}$$

$$\textbf{\textit{Client\_think}} \stackrel{def}{=} (think, r_t).\textbf{\textit{Client}}$$

The servers on the other hand first complete a *request* action followed by a *data* action in cooperation with the clients but at either stage they may perform a *break* action and enter a broken state in which a *reset* action is required before the server can be used again:

$$\textbf{\textit{Server}} \stackrel{def}{=} (request, r_r).\textbf{\textit{Server\_get}} + (break, r_b).\textbf{\textit{Server\_broken}}$$

$$\textbf{\textit{Server\_get}} \stackrel{def}{=} (data, r_d).\textbf{\textit{Server}} + (break, r_b).\textbf{\textit{Server\_broken}}$$

$$\textbf{\textit{Server\_broken}} \stackrel{def}{=} (reset, r_{rst}).\textbf{\textit{Server}}$$

The *request* and *data* actions are shared actions between the clients and servers in order to model the fact that clients must perform these actions by interacting with a server. The actions *timeout*, *think*, *break* and *reset*, on the other hand, are completed independently.

The PEPA model introduced above has n client components and m server components, each of which can be in one of three states, so the underlying CTMC of this simple model has the order of 3^{n+m} states. This exponential growth in the size of the state space for models of only modest description is known as the *state space explosion problem*. If state-space aggregation [20] is applied, the exponential growth still persists, only in the number of local derivative states rather than in the component population size [2]. For example, if $n = 200$, $m = 100$, there are of the order of $\binom{3+200-1}{200} \times \binom{3+100-1}{100} = 104, 570, 451$ aggregate states, still well beyond the capability of any explicit-state analysis method. This problem would of course be even more pronounced for more realistic and detailed models.

The natural representation of this situation as a GPEPA model would have the structure:

$$\textbf{\textit{CS}}(n, m) \stackrel{def}{=} \mathbf{Clients}\{\textbf{\textit{Client}}[n]\} \underset{L}{\bowtie} \mathbf{Servers}\{\textbf{\textit{Server}}[m]\} \tag{1.5}$$

where $C[k]$ extends to the fluid combinator $\wr\wr$. So the fluid components are ***Client***, ***Client_waiting***, ***Client_think***, ***Server***, ***Server_get*** and ***Server_broken***. That is, the fluid approximation will consist of six coupled differential equations counting the number of each of these fluid components active in the model. Alternatively, assuming that n is even, the following GPEPA model representation of the same standard PEPA model is also possible:

$$\mathbf{Clients}\{(\textbf{\textit{Client}} \parallel \textbf{\textit{Client}})[n/2]\} \underset{L}{\bowtie} \mathbf{Servers}\{\textbf{\textit{Server}}[m]\}$$

In this case, we consider each pair of clients as a single fluid component. So there will be twelve coupled differential equations since $|ds(\textbf{\textit{Client}} \parallel \textbf{\textit{Client}})| = 9$.

It is important to note that the combinator $\wr\wr$ has the same stochastic meaning as $\parallel$. However, as we have seen above, the two distinct combinators are necessary to resolve possible ambiguity in the case of component groups, which contain fluid components with their own internal parallelism. Indeed, the purpose of the additional level of model structure afforded by GPEPA models is to define the granularity at which the fluid approximation is performed, as will be described in Section 2.

We define the operational semantics of a GPEPA model to be identical to that of the equivalent standard PEPA model which is obtained syntactically by removing the group labels and replacing the $\wr\wr$ combinators with $\parallel$. We call this operation *flattening* and a formal *flattening function* is given in Definition 1.1. We may thus define the equivalent operational semantics on GPEPA models by composing the flattening function with the operational semantics of standard PEPA.

Formally, for any two GPEPA models, G_1 and G_2, we say $G_1 \xrightarrow{(\alpha, r)} G_2$ if and only if $\mathcal{F}(G_1) \xrightarrow{(\alpha, r)} \mathcal{F}(G_2)$, counting also the multiplicity of such transitions. In this way we may extend the definition of the set of derivative states $ds(G)$ to GPEPA models G. Further, we may define an *underlying CTMC* for a GPEPA model, which is trivially isomorphic to that of the corresponding equivalent standard PEPA model obtained through flattening.

Definition 1.1 (*Flattening Function*)**.** For any GPEPA model G, the corresponding standard PEPA model $\mathcal{F}(G)$, can be recovered from the grouped model, defined by: $\mathcal{F}(M_1 \underset{L}{\bowtie} M_2) := \mathcal{F}(M_1) \underset{L}{\bowtie} \mathcal{F}(M_2)$ and $\mathcal{F}(Y\{D\}) := \mathcal{F}'(D)$; where for component groups: $\mathcal{F}'(D_1 \wr\wr D_2) := \mathcal{F}'(D_1) \parallel \mathcal{F}'(D_2)$ and $\mathcal{F}'(P) := P$.

Also, we extend naturally to a GPEPA model G, the definitions of apparent rate $r_\alpha(G) := r_\alpha(\mathcal{F}(G))$ for $\alpha \in \mathcal{A}$, action set $\mathcal{A}(G) := \mathcal{A}(\mathcal{F}(G))$ and activity set $\mathcal{A}ct(G) := \mathcal{A}ct(\mathcal{F}(G))$.

1.1.4. Service level agreements and passage-time definition in Grouped PEPA

Service level agreements or SLAs form the basis of many industrial quality-of-service guarantees. In this paper, we tackle a class of SLAs which are specified in terms of passage-time quantiles, which can in general be described by the statement:

A sequence of events should occur within a time t, with a certain probability, p.

In order to validate an SLA of this type, we will, in general, require a passage-time distribution for the occurrence of the sequence of events within the system. To achieve this for models of realistic size, we will extract passage-time distributions from 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:

Global passage time *What is the probability that half of the clients have executed at least one think action by time t?*
Individual passage time *What is the probability that any individual server has completed a break action by time t?*

We will introduce these two classes of passage times in Section 3. In Section 4 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 $\boldsymbol{CS}(2n, n)$ 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 approximate 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.

2. Fluid analysis of GPEPA models

Fluid analysis captures the number of fluid components in a particular derivative state of a GPEPA model as the system evolves. The evolution of the fluid components is described by a set of ordinary differential equations, derived directly from the GPEPA model description. These differential equations are easy to solve numerically and provide a straightforward approach to analysing massive performance models. Fluid semantics for PEPA, first introduced by Hillston [3], have since been extended and developed in a number of different directions in the literature [21,22,2]. Furthermore, similar ideas have been applied in other stochastic process algebra [4,5] and stochastic Petri net [23] formalisms.

All of the material in this section is taken from [2]. In some cases, we give only informal explanations in this paper, such as the table below which enumerates key functions of a GPEPA model which will be used to generate the differential equations in Definition 2.4.

$\mathcal{G}(G)$	The set of all component group labels in the GPEPA model G, e.g. $\mathcal{G}(\boldsymbol{CS}(n, m)) = \{\textbf{Clients}, \textbf{Servers}\}$.
$\mathcal{B}(G, H)$	The set of all fluid component states in the component group of G which has group label H, e.g. $\mathcal{B}(\boldsymbol{CS}(n, m), \textbf{Clients}) = \{\boldsymbol{Client}, \boldsymbol{Client\_waiting}, \boldsymbol{Client\_think}\}$.
$\mathcal{B}(G)$	The set of all pairs of a component group label and a fluid component in the group specified by that label, e.g. $\mathcal{B}(\boldsymbol{CS}(n, m)) = \{(\textbf{Clients}, \boldsymbol{Client}), (\textbf{Clients}, \boldsymbol{Client\_waiting}), (\textbf{Clients}, \boldsymbol{Client\_think}), (\textbf{Servers}, \boldsymbol{Server}), (\textbf{Servers}, \boldsymbol{Server\_get}), (\textbf{Servers}, \boldsymbol{Server\_broken})\}$.
$\mathcal{N}(G)$	The number of all possible fluid component derivative states in each group of G, representing the number of approximating differential equations, $\mathcal{N}(G) = \lvert\mathcal{B}(G)\rvert$, e.g. $\mathcal{N}(\boldsymbol{CS}(n, m)) = 6$.
$\mathcal{S}(G, H)$	The size of the component group with label H. That is, the number of parallel components in the group, e.g. $\mathcal{S}(\boldsymbol{CS}(n, m), \textbf{Clients}) = n$.
$\mathcal{S}(G)$	The total size of all component groups in G, e.g. $\mathcal{S}(\boldsymbol{CS}(n, m)) = n + m$.

2.1. Deriving ODEs from GPEPA models

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

$\mathcal{R}_\alpha(G, N, H, P)$	The **component rate function** measures the local rate at which fluid components in state P in group H perform α-actions in the context of the cooperation within the wider grouped PEPA model G in aggregate state given by N.
$p_\alpha(P, Q)$	The **derivative weighting function** measures the probability that a fluid component P evolves to Q in one α-transition.
$r_\alpha(G, N)$	The **apparent rate function** measures the total rate of α being produced by GPEPA model G in aggregate state represented by N.

The quantities which will be subject to the fluid approximation are exposed formally through an aggregation of a GPEPA model's state space. Considering $\boldsymbol{CS}(n, m)$ again, we see there are $n \times m$ different ways the initial shared *request* action can be performed. This is because the *request* action involves exactly one ***Client*** and one ***Server*** component. Each of these transitions occurs at rate $\frac{1}{n}\frac{1}{m}\min(n, m)r_r$. The aggregation collects states together based on the number of fluid components in each

derivative state in each component group. In the case of $\boldsymbol{CS}(n, m)$, we might represent the initial aggregate state informally as "$n \times$ ***Client***, $0 \times$ ***Client_waiting***, $0 \times$ ***Client_think***, $m \times$ ***Server***, $0 \times$ ***Server_get*** and $0 \times$ ***Server_broken*** components". All of the $n \times m$ *request*-transitions would then become a single transition from this aggregate state to the aggregate state "$(n-1) \times$ ***Client***, $1 \times$ ***Client_waiting***, $0 \times$ ***Client_think***, $(m-1) \times$ ***Server***, $1 \times$ ***Server_get*** and $0 \times$ ***Server_broken*** components" at an aggregate rate of $\min(n, m)r_r$. The general extension of this aggregation process constructs an *underlying aggregated CTMC* from a given GPEPA model (as originally constructed for PEPA [20]).

In general and more formally, it has been shown [2, Theorem 2.12] that the underlying CTMC of a GPEPA model can always be aggregated according to the fluid 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 fluid component in each component group. Then each state of the underlying aggregated CTMC of a GPEPA model G can be uniquely determined by the model's initial state and a function $N \in \mathcal{B}(G) \to \mathbb{Z}_+$. This function counts the number of fluid components in each derivative state currently active in a given component group.

We may then define the *component rate function* for a GPEPA model G, which calculates the aggregate rate at which fluid components of type P within a component group H complete an action α in the aggregate state specified by N. This is needed to describe the rate of evolution of a component group from one derivative state to the next when constructing the differential equations from the model.

Definition 2.1 (*Component Rate Function*)**.** Let G be a GPEPA model. For $(H, P) \in \mathcal{B}(G)$, action type $\alpha \in \mathcal{A}$ and $N \in \mathcal{B}(G) \to \mathbb{Z}_+$ specifying the component counts, the component rate is $\mathcal{R}_\alpha(G, N, H, P)$, defined as:

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

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

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

This definition uses an alternate version of the apparent rate function, defined in terms of component counts, $N \in \mathcal{B}(G) \to \mathbb{Z}_+$. The apparent rate function, $r_\alpha(G, N)$, measures the total rate of α being produced by the whole GPEPA model G in the aggregate state represented by N.

Definition 2.2 (*Apparent Rate*)**.** Let G be a GPEPA model. Let $\alpha \in \mathcal{A}$ be an action type and $N \in \mathcal{B}(G) \to \mathbb{Z}_+$ specify the component counts. Then the apparent rate is $r_\alpha(G, N)$, defined as:

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

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

By way of example, consider the aggregate state of $\boldsymbol{CS}(n, m)$ represented by the function $N \in \mathcal{B}(\boldsymbol{CS}(n, m)) \to \mathbb{Z}_+$, defined by $N(\textbf{Clients}, \textbf{\textit{Client}}) = 1$, $N(\textbf{Clients}, \textbf{\textit{Client\_waiting}}) = 0$, $N(\textbf{Clients}, \textbf{\textit{Client\_think}}) = n - 1$, $N(\textbf{Servers}, \textbf{\textit{Server}}) = m$ and $N(\textbf{Servers}, \textbf{\textit{Server\_get}}) = N(\textbf{Servers}, \textbf{\textit{Server\_broken}}) = 0$. Then an example component rate function evaluation on this state is:

$$\mathcal{R}_{think}(\boldsymbol{CS}(n, m), N, \textbf{Clients}, \textbf{\textit{Client\_think}}) = (n - 1) \times r_t$$

In order to move towards the derivation of the fluid model (Definition 2.4) that will describe the evolution of a general GPEPA model over time, we need to introduce an explicit stochastic process that defines the state of the model. So for a GPEPA model G, let the integer-valued stochastic process $N_{H,P}(t)$ count the number of P-components active at a given time $t \in \mathbb{R}_+$ within the component group, H, for $(H, P) \in \mathcal{B}(G)$. In Definition 2.4, we will see how the process $N_{H,P}(t)$ can be approximated by a real-valued deterministic functions $v_{H,P}(t)$ by means of a system of ODEs.

In order to generate a set of ODEs for $v_{H,P}(t)$ that approximate $N_{H,P}(t)$, we require expressions for the total increment rate (part (1) of Eq. (2.2)) and the total decrement rate (part (2) of Eq. (2.2)) of a component P in the group H. We also need a counting process $N_t \in \mathcal{B}(G) \to \mathbb{Z}_+$ associated with $N_{H,P}(t)$ such that $N_t(H, P) = N_{H,P}(t)$ for all $(H, P) \in \mathcal{B}(G)$. N_t represents the aggregated CTMC state at time t.

The decrement rate attributed to an action α is the sum of the rates of all outgoing α-transitions from the current aggregated CTMC state which involve evolution of a P-component into some other component and can be shown to be [2, Theorem 2.15]:

$$\sum_{P \neq Q \in \mathcal{B}(G,H)} p_\alpha(P, Q) \mathcal{R}_\alpha(G, N_t, H, P)$$

where $p_\alpha(P, Q)$ is the *derivative weighting function* (defined below), the probability that a fluid component P doing an α-action transits to another specified fluid component Q. The increment rate of P in H attributed to an action α is the sum of the rates of all outgoing α-transitions from the current aggregated CTMC state which involve evolution into a P-component, given by [2, Theorem 2.15]:

$$\sum_{P \neq Q \in \mathcal{B}(G,H)} p_\alpha(Q, P)\, \mathcal{R}_\alpha(G, N_t, H, Q) \tag{2.1}$$

Definition 2.3 (*Derivative Weighting Function*)**.** Let P and Q be fluid 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 $r_\alpha(P) = 0$.

The rate of change of the number of components P in H is finally given by the increment rate (1) minus the decrement rate (2) in Eq. (2.2), over all actions α.

Definition 2.4 (*ODEs Associated With a GPEPA Model*)**.** Let G be a GPEPA 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:

$$\dot{v}_{H,P}(t) = \underbrace{\sum_{\alpha \in \mathcal{A}} \sum_{Q \in \mathcal{B}(G,H)} p_\alpha(Q, P)\, \mathcal{R}_\alpha(G, V_t, H, Q)}_{(1)} - \underbrace{\sum_{\alpha \in \mathcal{A}} \mathcal{R}_\alpha(G, V_t, H, P)}_{(2)} \tag{2.2}$$

where for $t \in \mathbb{R}_+$, $V_t \in \mathcal{B}(G) \to \mathbb{R}$ is a continuous counting process that produces a real-valued approximation of the component count for time t. Thus $V_t(H, P) := v_{H,P}(t)$ for all $(H, P) \in \mathcal{B}(G)$. Unless stated otherwise, the initial conditions $V_0 \in \mathcal{B}(G) \to \mathbb{R}_+$ are those naturally defined by the initial state of G.

By way of example, applying this definition to the model $\boldsymbol{CS}(n, m)$ results in the differential equations of Appendix A.1.

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 GPEPA 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 2.4 simply as $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$, where $\mathbf{v}(t) = (v_1(t), \ldots, v_{\mathcal{N}(G)}(t))^T \in \mathbb{R}_+^{\mathcal{N}(G)}$, so that, if i corresponds to (H, P), then $v_i(t) = v_{H,P}(t)$ for all $t \in \mathbb{R}_+$. 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 technical result in Appendix B.1 gives Lipschitz continuity of $\mathbf{f}$, thus guaranteeing the unique existence of a solution to the system of differential equations [e.g. 24]. It is also required for the proofs of convergence results in Section 4. Finally, the following straightforward result will prove to be fundamental in ensuring that passage-time approximations are comparable for a sequence of structurally-equivalent models.

Lemma 2.5. *Let G be a GPEPA 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, that is: $r_\alpha(G, \beta N) = \beta r_\alpha(G, N)$ and $\mathcal{R}_\alpha(G, \beta N, H, P) = \beta \mathcal{R}_\alpha(G, N, H, P)$ for all $(H, P) \in \mathcal{B}(G)$ and $N \in \mathcal{B}(G) \to \mathbb{R}_+$. □

It is known that $\mathbf{v}(t)$ can often be expected to approximate $\mathbb{E}[\mathbf{N}(t)]$ well [e.g. 2,18,25,8,4]. This fact will be exploited to construct some of the passage-time approximations of Section 3. Qualitative estimations of the accuracy of this approximation can be obtained by utilising the notion of so-called *switch points* as detailed in [18]. Switch points are coordinates in the ODE phase space where the fluid approximation can be expected to be at its worst. Fluid-approximation results from models for which the ODEs remain for long periods in such regions should then be treated with much more caution.

We may extend the ideas of this section to develop differential equations which approximate arbitrary higher-order moments of GPEPA model component counts [2]. Such approximations will help us still further in Section 3 to develop tighter global passage-time approximations. The additional differential equations which can be used to approximate higher-order moments are introduced in the next section. Similarly to the first moment approximation, the notion of switch points [18] can also be used to reason qualitatively about the approximation accuracy for higher moments.

2.2. Higher-moment differential equations from GPEPA models

For the sake of brevity we will not go into detail regarding how differential equations approximating higher-order moments of component counts are generated. The general approach is similar to *moment closure* techniques from biology and chemistry [e.g. 25–27]. In the case of GPEPA, we will give the definitions here and direct the reader towards [2] for the detailed derivation and further discussion.

In this section, we are interested in providing differential-equation based approximations to arbitrary moments of fluid component counts in GPEPA models. For a GPEPA model G, a general higher-order moment $\mathbb{E}[M(t)]$ can be specified by functions $\mathcal{M} \in \mathcal{B}(G) \to \mathbb{Z}_+$ such that $M(t) = \prod_{B \in \mathcal{B}(G)} N_B(t)^{\mathcal{M}(B)}$. We will write $v_{\mathcal{M}}(t)$ for its ODE approximation, the solution to the ODE given in the following definition.

Definition 2.6 (*Higher-moment ODEs Associated With a GPEPA Model*)**.** Let G be a GPEPA model. We define the evolution of the $v_{\mathcal{M}}(t)$ over time for $\mathcal{M} \in \mathcal{B}(G) \to \mathbb{Z}_+$ by the system of first-order coupled ODEs:

$$\begin{aligned} \dot{v}_{\mathcal{M}}(t) \quad = \quad & \sum_{(\mathcal{J}_-, \mathcal{J}_+, \alpha) \in \mathcal{J}(G)} \rho_\alpha(\mathcal{J}_-, \mathcal{J}_+) \\ & \times \sum_{K \in \mathcal{K}(\mathcal{J}_-, \mathcal{J}_+, \mathcal{M})} \left(\left[\prod_{B \in \mathcal{J}_- \setminus \mathcal{J}_+} (-1)^{\mathcal{M}(B) - K(B)} \prod_{B \in \mathcal{B}(G)} \binom{\mathcal{M}(B)}{K(B)} \right] \mathcal{R}_\alpha(G, V_t^*[K], \mathcal{J}_-) \right) \end{aligned} \tag{2.3}$$

For $t \in \mathbb{R}_+$, $V_t^*[K] \in \mathcal{B}(G) \to \mathbb{R}_+$ is given by: for all $B \in \mathcal{B}(G)$, $V_t^*[K](B) := v_{\mathcal{I}}(t)$ where $\mathcal{I}(B') = K(B')$ for all $B' \neq B$ and $\mathcal{I}(B) = K(B) + 1$. $\mathcal{K}(\mathcal{J}_-, \mathcal{J}_+, \mathcal{M})$ is the set of all elements $K \in \mathcal{B}(G) \to \mathbb{Z}_+$, such that $0 \leq K(B) \leq \mathcal{M}(B)$ for all $B \in \mathcal{J}_- \ominus \mathcal{J}_+$, and $K(B) = \mathcal{M}(B)$ for all $B \in \mathcal{B}(G) \setminus (\mathcal{J}_- \ominus \mathcal{J}_+)$, and $\sum_{B \in \mathcal{B}(G)} K(B) < \sum_{B \in \mathcal{B}(G)} \mathcal{M}(B)$.[1]

The initial condition for each moment approximation is that naturally implied by the initial state of G.

This definition requires a generalisation of the component rate function (Definition 2.1), the *joint component rate function*, which is the aggregate rate at which all of a given set of fluid components complete an action of a given action type in cooperation together. This is defined in terms of the *joint evolution set* which enumerates the possible ways in which fluid components can evolve together with a particular action type. Both definitions follow.

Definition 2.7 (*Joint Evolution Set*)**.** Let G be a GPEPA model. Then the joint evolution set is $\mathcal{J}(G)$, defined as follows.

$$\begin{aligned} \mathcal{J}(M_1 \underset{L}{\bowtie} M_2) := & \{(\mathcal{J}_-^1 \cup \mathcal{J}_-^2, \mathcal{J}_+^1 \cup \mathcal{J}_+^2, \alpha) : (\mathcal{J}_-^1, \mathcal{J}_+^1, \alpha) \in \mathcal{J}(M_1), (\mathcal{J}_-^2, \mathcal{J}_+^2, \alpha) \in \mathcal{J}(M_2), \alpha \in L\} \\ & \cup \{(\mathcal{J}_-, \mathcal{J}_+, \alpha) : (\mathcal{J}_-, \mathcal{J}_+, \alpha) \in \mathcal{J}(M_1), \alpha \notin L\} \\ & \cup \{(\mathcal{J}_-, \mathcal{J}_+, \alpha) : (\mathcal{J}_-, \mathcal{J}_+, \alpha) \in \mathcal{J}(M_2), \alpha \notin L\} \\ \mathcal{J}(H\{D\}) := & \bigcup_{\substack{P \xrightarrow{(\alpha, \cdot)} Q \\ P, Q \in \mathcal{B}(G,H)}} \{\{(H, P)\}, \{(H, Q)\}, \alpha\} \end{aligned}$$

Definition 2.8 (*Joint Component Rate Function*)**.** Let G be a GPEPA model. Let $\mathcal{J} \subseteq \mathcal{B}(G)$ be non-empty. Let $\alpha \in \mathcal{A}$ and $N \in \mathcal{B}(G) \to \mathbb{Z}_+$. Then the joint component rate is $\mathcal{R}_\alpha(G, N, \mathcal{J})$, defined as follows.

$$\mathcal{R}_\alpha(M_1 \underset{L}{\bowtie} M_2, N, \mathcal{J}) := \begin{cases} \frac{\mathcal{R}_\alpha(M_1, N, \mathcal{J}_1)}{r_\alpha(M_1, N)} \frac{\mathcal{R}_\alpha(M_2, N, \mathcal{J}_2)}{r_\alpha(M_2, N)} \min(r_\alpha(M_1, N), r_\alpha(M_2, N)) & \text{if } \alpha \in L, \mathcal{J} \not\subseteq \mathcal{B}(M_j) \text{ for } j = 1 \text{ and } 2 \\ \mathcal{R}_\alpha(M_j, N, \mathcal{J}) & \text{if } \alpha \notin L, \mathcal{J} \subseteq \mathcal{B}(M_j) \text{ for } j = 1 \text{ or } 2 \\ 0 & \text{otherwise} \end{cases}$$

In the first line of the definition, we define $\mathcal{J}_1$ and $\mathcal{J}_2$ to be the unique partition of $\mathcal{J}$ such that $\mathcal{J}_1 \subseteq \mathcal{B}(M_1)$ and $\mathcal{J}_2 \subseteq \mathcal{B}(M_2)$. As before, terms with zeros in the denominator are defined as zero.

$$\mathcal{R}_\alpha(H\{D\}, N, \mathcal{J}) := \begin{cases} N(H, P)\, r_\alpha(P) & \text{if } \mathcal{J} = \{(H, P)\} \\ 0 & \text{otherwise} \end{cases}$$

Finally, we require a generalised version of the derivative weighting function (Definition 2.3), the *joint derivative weighting function*, which for a given set of fluid components, computes the probability that after a joint α-action they transit together to another given set of fluid components.

Definition 2.9 (*Joint Derivative Weighting Function*)**.** Let G be a GPEPA model and let $(\mathcal{J}_-, \mathcal{J}_+, \alpha) \in \mathcal{J}(G)$. Then $\rho_\alpha(\mathcal{J}_-, \mathcal{J}_+)$ is the joint derivative weighting function, defined as follows.

$$\rho_\alpha(\mathcal{J}_-, \mathcal{J}_+) := \prod_{(H,P) \in \mathcal{J}_-} p_\alpha(P, Q_{H, \mathcal{J}_+})$$

where $Q_{H, \mathcal{J}_+}$ is defined as the unique fluid component Q, such that $(H, Q) \in \mathcal{J}_+$. That it will exist and be unique is guaranteed by the fact that $(\mathcal{J}_-, \mathcal{J}_+, \alpha) \in \mathcal{J}(G)$.

[1] Where $A \ominus B := (A \cup B) \setminus (A \cap B)$ is the symmetric difference of the two sets A and B.

3. 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. [15] 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 [28,29] where absorbing modifications are made to make passage-time measures more explicit.

For example, in the case of the $\boldsymbol{CS}(n, m)$ model of Eq. (1.5), we may be interested in how long it takes for some proportion of the initial n clients to complete their first cycle (consisting of at least one *request*-, *data*- and *think*-action, perhaps interrupted by *timeout*-actions). 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 ***Client_think*** fluid component and introduce three new fluid components, ***Client***′, ***Client_waiting***′ and ***Client_think***′, as follows:

$$\begin{aligned}
\boldsymbol{Client} &\stackrel{def}{=} (request, r_r).\boldsymbol{Client_waiting} & \boldsymbol{Client}' &\stackrel{def}{=} (request, r_r).\boldsymbol{Client_waiting}'\\
\boldsymbol{Client_waiting} &\stackrel{def}{=} (data, r_d).\boldsymbol{Client_think} & \boldsymbol{Client_waiting}' &\stackrel{def}{=} (data, r_d).\boldsymbol{Client_think}'\\
&\quad + (timeout, r_{tmt}).\boldsymbol{Client} & &\quad + (timeout, r_{tmt}).\boldsymbol{Client}'\\
\boldsymbol{Client_think} &\stackrel{def}{=} (think, r_t).\boldsymbol{Client}' & \boldsymbol{Client_think}' &\stackrel{def}{=} (think, r_t).\boldsymbol{Client}'
\end{aligned}$$

Call the resulting model $\boldsymbol{CS}'(n, m) \stackrel{def}{=} \mathbf{Clients}\{\boldsymbol{Client}[n]\} \underset{L}{\bowtie} \mathbf{Servers}\{\boldsymbol{Server}[m]\}$. These additional states will allow us to distinguish between components which have completed a cycle and ones that have not. Therefore, we are now in a position to express the random variable we are interested in as the time to extinction of the specified number of the ***Client***, ***Client_waiting*** and ***Client_think*** components in the modified model. 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 client to complete *any* number of cycles. We will shortly show how the differential equations obtained by applying Definition 2.4 to this modified model 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.

3.1. Global passage times

We consider again instances of the model introduced above, with even numbers of clients, that is, $\boldsymbol{CS}'(2n, m)$. Consider the passage-time quantity for half (or n) of the clients to complete their initial cycle. As mentioned above, the fluid semantics of Section 2.1 can be applied to this model, yielding the system of 9 ODEs given in Appendix A.2. In contrast to the case of the unmodified model, $\boldsymbol{CS}(2n, m)$, 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 $v_{\boldsymbol{C}}(t) + v_{\boldsymbol{Cw}}(t) + v_{\boldsymbol{Ct}}(t)$ (or $v_{\boldsymbol{C}'}(t) + v_{\boldsymbol{Cw}'}(t) + v_{\boldsymbol{Ct}'}(t)$) and computing the time t at which it reaches the value n.[2] We will 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 GPEPA model together with an absorbing subset of its aggregated state space, specified by a particular subset of fluid components and a target count for these fluid components to reach. The passage time is then the time taken for this to occur. The formal definition follows.

Definition 3.1 (*Global Passage Time*)**.** Let G be a GPEPA model, $\mathcal{C} \subseteq \mathcal{B}(G)$ be a subset of fluid components and $C \in \mathbb{Z}_+$ represent the target component count. Define the global passage-time random variable, $\sigma := \inf\{t \in \mathbb{R}_+ : \sum_{B\in\mathcal{C}} N_B(t) \leq C\}$, where whenever $t > \sigma$, $\sum_{B\in\mathcal{C}} N_B(t) \leq C$, that is, the target states must be absorbing and the passage is timed starting from the initial state of G.

This definition can be used to describe the example passage time for half of the clients to complete their first cycle by letting $G = \boldsymbol{CS}'(2n, m)$, $\mathcal{C} = \{\boldsymbol{Client}, \boldsymbol{Client\_waiting}, \boldsymbol{Client\_think}\}$ and $C = n$.

[2] Throughout this paper, we will often adopt obvious shorthands such as $v_{\boldsymbol{C}}(t)$ for $v_{\mathbf{Clients},\boldsymbol{Client}}(t)$ and $v_{\boldsymbol{Sb}}(t)$ for $v_{\mathbf{Servers},\boldsymbol{Server\_broken}}(t)$.

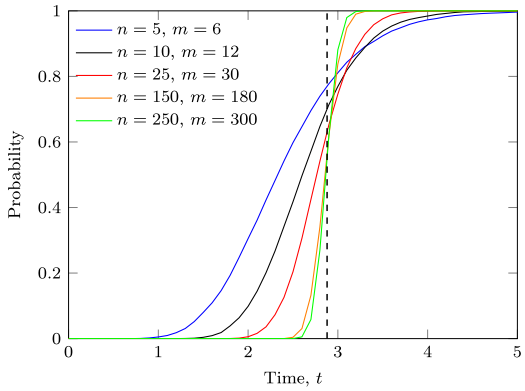


Fig. 2. Passage-time CDFs for half of the clients to complete a cycle computed by stochastic simulation compared with the ODE point-mass approximation (the dashed vertical line) for the model $\boldsymbol{CS}(2n, m)$.

3.1.1. Point-mass approximation

The most straightforward approach to approximating the passage time mentioned above would be to compute the time t at which the quantity $v_{\boldsymbol{C}}(t) + v_{\boldsymbol{Cw}}(t) + v_{\boldsymbol{Ct}}(t)$ reaches n. Fig. 2 shows cumulative distribution functions computed using traditional methods for this passage-time random variable, with model rates set to $r_r = 2.0$, $r_{tmt} = 0.3$, $r_t = 0.5$, $r_b = 0.05$, $r_d = 2.0$, $r_{rst} = 1.0$. In each case, we increase the number of clients and there are always three fifths as many servers as there are clients. Maintaining this ratio ensures that the point-mass approximation for each of these passage times is actually the same (Lemma 2.5), represented by the dashed vertical line in the figure. The probability density functions converge to the point mass as the component populations increase.

For a general global passage time specified for some GPEPA model G by $\mathcal{C} \subseteq \mathcal{B}(G)$ and $C \in \mathbb{Z}_+$, as in Definition 3.1, the point-mass approximation is defined simply as $\inf\{t \in \mathbb{R}_+ : \sum_{B\in\mathcal{C}} v_B(t) \leq C\}$. In Section 4.3.1, we will show that the limiting result depicted in Fig. 2 holds in general.

3.1.2. Upper and lower CDF approximations

For smaller populations sizes, many of the passage-time distributions depicted in Fig. 2 have a significant level of variability. In such cases, a deterministic passage-time approximation does not capture an accurate picture. In this section, we show how we might address this by introducing what we term *upper* and *lower* approximations to global passage-time CDFs. These are ODE-computed approximations to theoretically exact *upper* and *lower* CDF bounds. Often the approximation itself is very accurate but this can be balanced by bounds which are not always as tight as we might ideally like.

In the next section we will introduce the simplest such approximations computed using only the first-moment ODE approximations for GPEPA models which were defined in Section 2.1. Then we will develop tighter approximations utilising the higher-moment approximations of Section 2.2.

3.1.3. First-moment CDF approximations

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

$$\mathbb{P}\{X \geq a\} \leq \frac{\mathbb{E}[X]}{a} \tag{3.1}$$

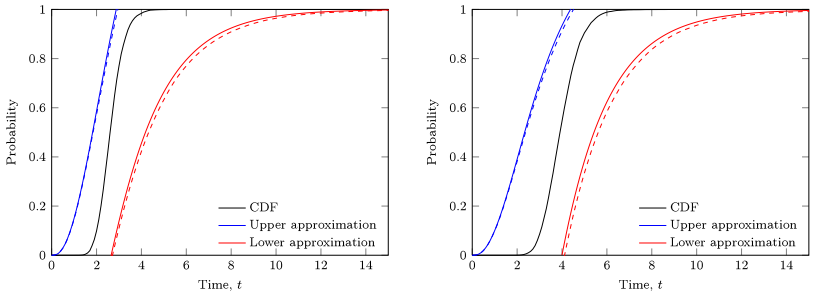
In order to exploit this, we can use the system of ODEs defined in Section 2.1 to approximate the component-count expectations.

Consider again the model $\boldsymbol{CS}'(2n, m)$ and the passage time for n of the clients 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:

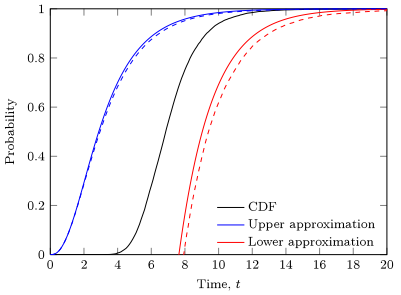
$$\begin{aligned}\mathbb{P}\{\sigma \leq t\} &= \mathbb{P}\{N_{\boldsymbol{C}'}(t) + N_{\boldsymbol{Cw}'}(t) + N_{\boldsymbol{Ct}'}(t) \geq n\} \leq \frac{\mathbb{E}[N_{\boldsymbol{C}'}(t)] + \mathbb{E}[N_{\boldsymbol{Cw}'}(t)] + \mathbb{E}[N_{\boldsymbol{Ct}'}(t)]}{n} \\ &= 1 - \mathbb{P}\{N_{\boldsymbol{C}}(t) + N_{\boldsymbol{Cw}}(t) + N_{\boldsymbol{Ct}}(t) \geq n+1\} \geq 1 - \frac{\mathbb{E}[N_{\boldsymbol{C}}(t)] + \mathbb{E}[N_{\boldsymbol{Cw}}(t)] + \mathbb{E}[N_{\boldsymbol{Ct}}(t)]}{n+1}\end{aligned} \tag{3.2}$$

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.

Fig. 3 gives the results of applying this approach for three different passage-time random variables. In all three cases we note that the ODE approximation to the theoretical bound is generally very accurate. Where there is a larger discrepancy



(a) Half of the clients. (b) Three quarters of the clients.



(c) All of the clients.

Fig. 3. Passage-time CDFs for varying proportions of the clients to complete a cycle computed by stochastic simulation compared with the ODE-derived upper and lower approximations from Markov's inequality for the model ***CS***(20, 12). The dashed line represents the actual theoretical bound again computed by stochastic simulation. Rates: $r_r = 2.0, r_{tmt} = 0.3, r_t = 0.5, r_b = 0.05, r_d = 2.0, r_{rst} = 1.0$.

between the ODEs and the simulated means, the error results from the approximation used to derive a closed system of ODEs in Definition 2.4 and is concentrated in time intervals containing so-called *switch points*, as discussed in [2,18]. The technique from [18] can detect the time intervals where these switch points exist and can suggest a suitable change in system parameters to obtain more accurate ODEs.

We would certainly expect a more accurate result than the point-mass approximation depicted in Fig. 2 since that required the entire distribution to concentrate around a point mass. In this section, however, we require only a convergence of expectations, therefore, it would seem that the approximations of this section may be more useful for smaller component populations than the point-mass approximation. This is an improvement paid for by the fact that only bounds on the CDF can be obtained. However, a lower bound on a CDF is sufficient to verify satisfaction of an SLA specified in terms of passage-time quantiles.

It is also pertinent to note that as the population size increases, these bounds can become quite loose, so at some point, it is certainly likely to be advantageous to switch to the point-mass approximation. Indeed, when only first moments are considered, we would not expect that the bounds would become tighter in the limit of large populations since no measure of variability is considered.

Another interesting point to note in Fig. 3 is that the relative tightness of the more useful lower approximation appears to be increasing in the higher and arguably more useful quantiles as the proportion being timed increases, whereas the upper approximation becomes looser everywhere. To see why this is to be expected let σ_a be the passage-time random variable for a of the clients to complete their first cycle and then note that in this case, we may phrase Eq. (3.2) as:

$$\mathbb{E}[N_{\mathbf{C}}(t)] + \mathbb{E}[N_{\mathbf{Cw}}(t)] + \mathbb{E}[N_{\mathbf{Ct}}(t)] \geq \mathbb{P}\{\sigma_a > t\} \times (2n - a + 1)$$

This inequality can also be derived by considering the events $\{\sigma_a > t\}$ and $\{\sigma_a \leq t\}$ and the smallest possible values of $N_{\mathbf{C}}(t) + N_{\mathbf{Cw}}(t) + N_{\mathbf{Ct}}(t)$ on each event. If $\sigma_a \leq t$, the passage has completed and without extra information, we cannot say anything more than the trivial statement $N_{\mathbf{C}}(t) + N_{\mathbf{Cw}}(t) + N_{\mathbf{Ct}}(t) \geq 0$. In general, after a passage has completed, $N_{\mathbf{C}}(t) + N_{\mathbf{Cw}}(t) + N_{\mathbf{Ct}}(t)$ will not actually be zero and could in fact be as large as $2n - a$. However as a approaches $2n$ the potential for this discrepancy decreases hence we would expect the approximation to become tighter as we observe in Fig. 3. A similar argument can be made for the increasing looseness of the upper approximation.

First-moment CDF approximations for general global passage times. We now show how to bound global passage-time random variables in general. Let σ be a global passage-time random variable specified for some GPEPA model G by $\mathcal{C} \in \mathcal{B}(G)$ and $C \in \mathbb{Z}_+$, as in Definition 3.1. We may then compute, writing $\mathcal{H} := \{H : (H, P) \in \mathcal{C}\}$ for the component groups involved in the specification of the global passage time:

$$\begin{aligned}\mathbb{P}\{\sigma \leq t\} &= \mathbb{P}\left\{\sum_{B\in\mathcal{C}} N_B(t) \leq C\right\}\\ &= \mathbb{P}\left\{\sum_{H_i\in\mathcal{H}}\sum_{(H_i,P_j)\in\mathcal{C}} N_{H_i,P_j}(t) \leq C\right\}\\ &= \mathbb{P}\left\{\sum_{H_i\in\mathcal{H}}\left(\mathcal{S}(G,H_i) - \sum_{(H_i,P_j)\in\mathcal{C}} N_{H_i,P_j}(t)\right) \geq \sum_{H_i\in\mathcal{H}}\mathcal{S}(G,H_i) - C\right\}\end{aligned}$$

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

$$\mathbb{P}\{\sigma \leq t\} \leq \frac{\sum_{H_i\in\mathcal{H}}(\mathcal{S}(G,H_i) - \sum_{(H_i,P_j)\in\mathcal{C}}\mathbb{E}[N_{H_i,P_j}(t)])}{\sum_{H_i\in\mathcal{H}}\mathcal{S}(G,H_i) - C} \tag{3.3}$$

Working in the other direction, we have:

$$\mathbb{P}\{\sigma \leq t\} = 1 - \mathbb{P}\left\{\sum_{B\in\mathcal{C}} N_B(t) > C\right\} = 1 - \mathbb{P}\left\{\sum_{B\in\mathcal{C}} N_B(t) \geq C+1\right\} \tag{3.4}$$

Applying Markov's inequality directly, we obtain:

$$\mathbb{P}\{\sigma \leq t\} \geq 1 - \frac{1}{C+1}\sum_{(H,P)\in\mathcal{C}}\mathbb{E}[N_{H,P}(t)] \tag{3.5}$$

The approximation $\mathbb{E}[\mathbf{N}(t)] \approx \mathbf{v}(t)$ can then be applied directly in either case to provide the bound estimates.

3.1.4. Higher-moment CDF approximations

In this section, we show how the techniques of the last section can be improved by exploiting the differential-equation approximations to higher-order moments as constructed in Section 2.2.

We will begin by replacing the use of Markov's inequality in the previous section with Chebyshev's inequality. If X is an arbitrary random variable, $t > 0$ and $q \neq 0$, Chebyshev's inequality says [e.g. 30, Theorem A.113, Page 492]:

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

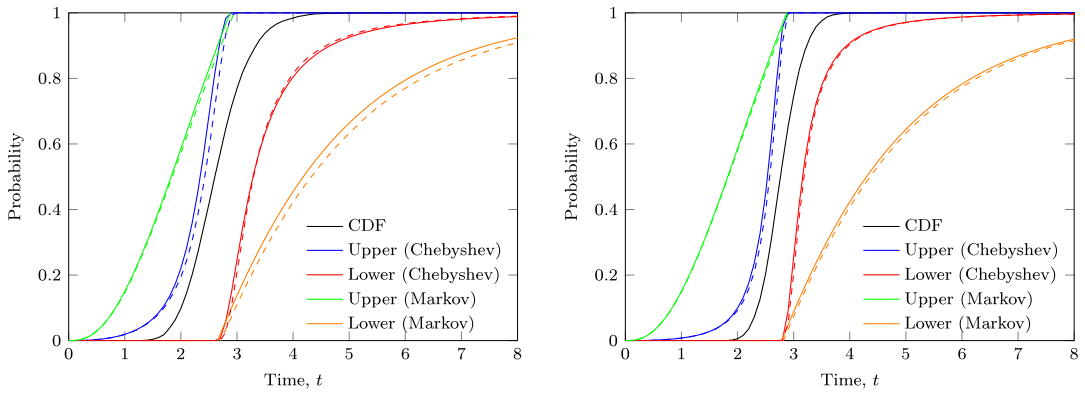
For now, we will consider the case $q = 2$ which introduces variance information. This should result in bounds that are tighter than those of the last section. In the case $q = 2$, it is a straightforward application of the Cauchy–Schwarz inequality to derive a one-sided refinement of this, often known as the Chebyshev–Cantelli inequality, which says that for $b > 0$:

$$\mathbb{P}\{X - \mathbb{E}[X] \geq b\} \leq \frac{\mathrm{Var}[X]}{\mathrm{Var}[X] + b^2} \qquad \mathbb{P}\{\mathbb{E}[X] - X \geq b\} \leq \frac{\mathrm{Var}[X]}{\mathrm{Var}[X] + b^2}$$

As before, let σ be a global passage-time random variable specified for some GPEPA model G by $\mathcal{C} \in \mathcal{B}(G)$ and $C \in \mathbb{Z}_+$, as in Definition 3.1. Then we have:

$$\begin{aligned}\mathbb{P}\{\sigma \leq t\} = \mathbb{P}\left\{\sum_{B\in\mathcal{C}} N_B(t) \leq C\right\} &= \mathbb{P}\left\{\sum_{B\in\mathcal{C}}(\mathbb{E}[N_B(t)] - N_B(t)) \geq \sum_{B\in\mathcal{C}}\mathbb{E}[N_B(t)] - C\right\}\\ &\leq \frac{\mathrm{Var}\left[\sum_{B\in\mathcal{C}} N_B(t)\right]}{\mathrm{Var}\left[\sum_{B\in\mathcal{C}} N_B(t)\right] + (\mathbb{E}\left[\sum_{B\in\mathcal{C}} N_B(t)\right] - C)^2}\end{aligned} \tag{3.6}$$

where the inequality is valid when $\mathbb{E}\left[\sum_{B\in\mathcal{C}} N_B(t)\right] - C > 0$. Similarly, working in the other direction, we may obtain:



(a) $n = 10, m = 12$. (b) $n = 25, m = 30$.

Fig. 4. Passage-time CDFs for half of the clients to complete a cycle computed by stochastic simulation compared with the ODE-derived upper and lower approximations from Chebyshev's inequality for the model $\boldsymbol{CS}(2n, m)$. The dashed line represents the actual theoretical bound again computed by stochastic simulation. Rates: $r_r = 2.0, r_{tmt} = 0.3, r_t = 0.5, r_b = 0.05, r_d = 2.0, r_{rst} = 1.0$.

$$\mathbb{P}\{\sigma \leq t\} = 1 - \mathbb{P}\left\{\sum_{B\in\mathcal{C}} N_B(t) \geq C + 1\right\} \tag{3.7}$$

$$= 1 - \mathbb{P}\left\{\sum_{B\in\mathcal{C}} (N_B(t) - \mathbb{E}[N_B(t)]) \geq C + 1 - \sum_{B\in\mathcal{C}} \mathbb{E}[N_B(t)]\right\}$$

$$\geq 1 - \frac{\mathrm{Var}\left[\sum_{B\in\mathcal{C}} N_B(t)\right]}{\mathrm{Var}\left[\sum_{B\in\mathcal{C}} N_B(t)\right] + (\mathbb{E}[\sum_{B\in\mathcal{C}} N_B(t)] - C - 1)^2} \tag{3.8}$$

where the inequality is valid when $C+1-\mathbb{E}[\sum_{B\in\mathcal{C}} N_B(t)] > 0$. We may then apply the differential-equation approximations to first- and second-order moments defined in Section 2 to yield approximations to these bounds. Fig. 4 shows how these bounds can substantially improve on those obtained using Markov's inequality.[3] We also observe that they appear to increase substantially in tightness as the component populations increase.

Both of the first- and second-order CDF approximations above are obtained by addressing the general problem in probability theory of retrieving a distribution from a selection of its moments, the so-called *reduced moment problem*. Application of the Markov and Chebyshev inequalities results in the simple arithmetic expressions for the lower and upper CDF approximations, Eqs. (3.5) and (3.8); and Eqs. (3.3) and (3.6), respectively. A more advanced technique is described by Tari et al. [32] that can utilise moments of orders higher than 2. We apply this technique later to obtain tighter approximations in Fig. 7(a) in Section 5 for our worked example model. There, we will show CDF approximations based on the first 4 moments, where for the first time the input moments are those generated by fluid techniques.

Finally, we note that for global passage times timing the complete extinction of a population (such as that depicted in Fig. 3(c)), the point-mass approximation of Section 3.1.1 cannot be applied directly since the relevant ODE quantity never actually reaches zero. In these cases, then, we would advise the practitioner to use the upper and lower CDF approximation techniques of this section.

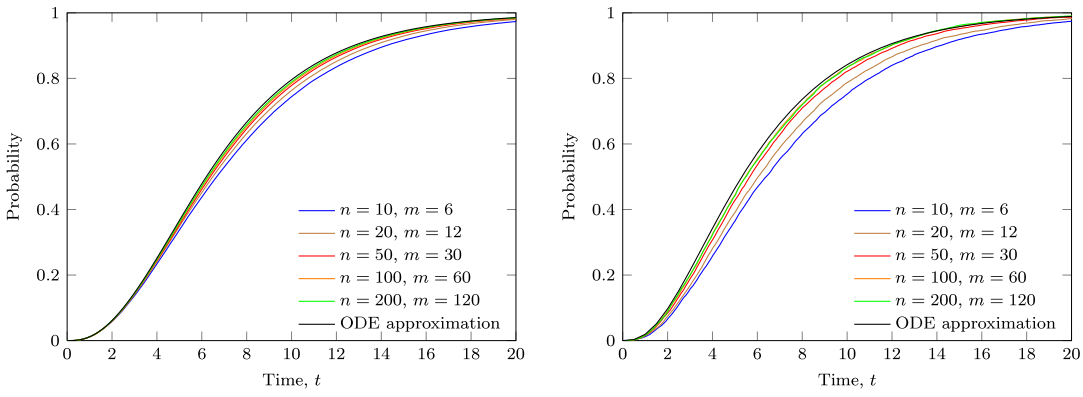
3.2. Individual passage times

In this section we will consider a second kind of passage-time measurement, the *individual passage time*. These are marginal passage times for individuals (fluid components) which are part of a large population of similar fluid components. We will show how the entire cumulative distribution function of such random variables can be approximated directly. This builds on approaches [12,13,10] which are able to obtain *average* individual passage-time measures by combining ODE-based techniques and Little's Law [11].

By means of introduction, consider again the model $\boldsymbol{CS}'(n, m)$ and in contrast to the last section, assume now that we are interested in how long it takes for one of the initial n clients to complete their first cycle. Since all members of the **Clients** component group are identically distributed it makes sense to speak of this passage-time measurement independent of any specific individual. Let $C_j(t) \in ds(\boldsymbol{Client})$ for $1 \leq j \leq n$ be the stochastic process which tracks the state of the jth individual in the ***Client*** component group. We wish to evaluate at time $t \in \mathbb{R}_+$ and for any $1 \leq k \leq n$, the cumulative distribution function:

$$\mathbb{P}\{C_k(t) \in \{\boldsymbol{Client}', \boldsymbol{Client\_waiting}', \boldsymbol{Client\_think}'\}\}$$

[3] We have made the full system of 54 ODEs approximating second-order moments for the model $\boldsymbol{CS}'(n, m)$ available online [31].



(a) Transient. (b) Steady state.

Fig. 5. Passage-time CDFs for a single client to complete a cycle computed by stochastic simulation compared with the ODE-derived approximation for the model $\boldsymbol{CS}(n, m)$. Rates: $r_r = 1.0, r_{tmt} = 0.1, r_t = 0.27, r_b = 0.1, r_d = 0.7, r_{rst} = 0.5$.

That is, the probability that by time t, a specific individual client has completed its first cycle. Now note that:

$$\mathbb{E}[N_{\boldsymbol{C}'}(t) + N_{\boldsymbol{Cw}'}(t) + N_{\boldsymbol{Ct}'}(t)] = \sum_{j=1}^{n} \mathbb{P}\{C_j(t) \in \{\boldsymbol{Client}', \boldsymbol{Client\_waiting}', \boldsymbol{Client\_think}'\}\}$$

So we compute the quantity of interest as:

$$\mathbb{P}\{C_k(t) \in \{\boldsymbol{Client}', \boldsymbol{Client\_waiting}', \boldsymbol{Client\_think}'\}\} = \frac{1}{n}(\mathbb{E}[N_{\boldsymbol{C}'}(t)] + \mathbb{E}[N_{\boldsymbol{Cw}'}(t)] + \mathbb{E}[N_{\boldsymbol{Ct}'}(t)])$$

If we again make the approximation $\mathbb{E}[\mathbf{N}(t)] \approx \mathbf{v}(t)$ this provides the route to the ODE approximation of the entire cumulative distribution function of the individual passage time.

Fig. 5(a) shows cumulative distribution functions computed using traditional methods for the passage-time random variable discussed above. In each case, we increase the number of clients and there are always three fifths as many servers as there are clients. This ensures, similarly to the previous section, that the differential-equation approximation to the distribution for each of these passage times is identical (Lemma 2.5). The cumulative distribution functions converge to the ODE approximation as the component populations increase. This result is proven in general in Section 4.3.2.

We call this, the simplest type of individual passage time, a *transient individual passage time*. The general classification of such passage times now follows.

Definition 3.2 (*Transient Individual Passage Time*)**.** Let G be a GPEPA model and $(H, P) \in \mathcal{B}(G)$ be the fluid component under observation and assume further that all fluid components in group H are initially P components. Let $\mathcal{T} \subseteq ds(P)$ be the set of target states, which is absorbing in the sense that $ds(\mathcal{T}) \subseteq \mathcal{T}$.

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

This definition can be used to describe the example passage time introduced earlier with $G := \boldsymbol{CS}'(n, m)$, $P := \boldsymbol{Client}$ and $\mathcal{T} := \{\boldsymbol{Client}', \boldsymbol{Client\_waiting}', \boldsymbol{Client\_think}'\}$.

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

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

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.

3.2.1. Steady-state individual passage times

In contrast to timing a component from a fixed initial state of the model, we will often be interested in measuring the time taken for a fluid component to move from one of a set of designated *start states* to a *target state* assuming that the model is in steady-state when the measurement is started. In order to compute such passage times, we will consider here two GPEPA models: one that is used to compute the model's steady-state (or stationary) distribution; and one where the fluid component under observation has a set of local derivative states which are absorbing (the target states) that is used to measure the passage time.

Consider again the model $\boldsymbol{CS}(n, m)$. Consider the same passage time as in the previous section: the time taken for a client to complete one cycle, but measured from the moment when the model is operating in the steady-state regime. As

mentioned above, we will need to use both the unmodified client–server GPEPA model $\boldsymbol{CS}(n, m)$ and the version with the absorbing modification $\boldsymbol{CS}'(n, m)$. The model $\boldsymbol{CS}(n, m)$ will be used to compute the stationary distribution with which we will initialise the model $\boldsymbol{CS}'(n, m)$ in order to compute the actual passage-time distribution.

The first stage in the computation of passage times of this form is to approximate the stationary expectations of the fluid component counts of $\boldsymbol{CS}(n, m)$ whose approximating system of ODEs was given in Section 2. Since we are interested in the stationary expectations, it makes sense to take the limit of the ODE solutions as $t \to \infty$. To compute these quantities, one can either integrate the ODEs numerically for a sufficiently large period of time or attempt to find a unique meaningful fixed point of the ODEs[4] by solving the system of algebraic equations obtained when the right-hand sides of the ODEs are equated with zero. We will write for example $v_{\boldsymbol{C}}$ for the ODE approximation to the stationary expectation $N_{\boldsymbol{C}} := \lim_{t\to\infty} \mathbb{E}[N_{\boldsymbol{C}}(t)]$ and so on for the other fluid components in $\boldsymbol{CS}(n, m)$. We will write component counts in the model $\boldsymbol{CS}'(n, m)$ with primes, for example, $N'_{\boldsymbol{C}}(t)$ and $N'_{\boldsymbol{C}'}(t)$; and similarly for their first-moment ODE approximations $v'_{\boldsymbol{C}}(t)$ and $v'_{\boldsymbol{C}'}(t)$, respectively.

The approximation of the passage-time CDF will then be obtained by solving the system of ODEs corresponding to $\boldsymbol{CS}'(n, m)$, initialised with the stationary expectations obtained above. To see how this might work, we consider the model $\boldsymbol{CS}(n, m)$ evolving in its stationary regime and then at some fixed time its evolution is switched to that of $\boldsymbol{CS}'(n, m)$, initialised by the current state of $\boldsymbol{CS}(n, m)$. The idea is then similar to that of the last section: by timing how long it takes for a client in the ***Client*** state to reach one of the target states in $\mathcal{T} := \{\boldsymbol{Client}', \boldsymbol{Client\_waiting}', \boldsymbol{Client\_think}'\}$ we can measure the duration of the passage.

To see how we proceed, define $C_j(t) \in ds(\boldsymbol{Client})$ to be the stochastic process, which tracks the state of the jth fluid component in the group **Clients**, as the model $\boldsymbol{CS}(n, m)$ evolves. Denote time 0 as the arbitrary fixed time at which the evolution of the model switches to $\boldsymbol{CS}'(n, m)$, initialised by the state of $\boldsymbol{CS}(n, m)$. Then the quantity we wish to compute is:

$$\mathbb{P}\{C_j(t) \in \mathcal{T} \mid C_j(0) = \boldsymbol{Client}\} \tag{3.9}$$

We can compute $\mathbb{P}\{C_j(0) = \boldsymbol{Client}\} = \frac{1}{n}\mathbb{E}[N_{\boldsymbol{C}}(0)] = \frac{1}{n}N_{\boldsymbol{C}}$. Therefore, in order to compute the conditional expression of Eq. (3.9), it suffices to find:

$$\mathbb{P}(\{C_j(t) \in \mathcal{T}\} \cap \{C_j(0) = \boldsymbol{Client}\})$$

It turns out that we can gain access to this quantity also in terms of the component-count expectations of $\boldsymbol{CS}'(n, m)$. However, this requires that we initialise the fluid component counts in $\boldsymbol{CS}'(n, m)$ slightly differently at time 0 than may initially be expected:

$$\begin{array}{lll} N'_{\boldsymbol{C}}(0) := N_{\boldsymbol{C}}(0) & N'_{\boldsymbol{C}'}(0) := 0 & N'_{\boldsymbol{S}}(0) := N_{\boldsymbol{S}}(0) \\ N'_{\boldsymbol{Cw}}(0) := 0 & N'_{\boldsymbol{Cw}'}(0) := N_{\boldsymbol{Cw}}(0) & N'_{\boldsymbol{Sg}}(0) := N_{\boldsymbol{Sg}}(0) \\ N'_{\boldsymbol{Ct}}(0) := 0 & N'_{\boldsymbol{Ct}'}(0) := N_{\boldsymbol{Ct}}(0) & N'_{\boldsymbol{Sb}}(0) := N_{\boldsymbol{Sb}}(0) \end{array}$$

Specifically, we do not initialise $N'_{\boldsymbol{Cw}}(0)$ with $N_{\boldsymbol{Cw}}(0)$, or $N'_{\boldsymbol{Ct}}(0)$ with $N_{\boldsymbol{Ct}}(0)$, but instead we place such client component counts in their equivalent absorbing states in $\boldsymbol{CS}'(n, m)$. This does not affect the value of Eq. (3.9) but does allow us to single out the client fluid components which were at the start of a passage (in state ***Client***) at time 0 and track their evolution through to time t. In this way we now have that[5]:

$$\sum_{Q\in\mathcal{T}} \left(N'_Q(t) - N'_Q(0)\right) = \sum_{j=1}^{n} \mathbf{1}_{\{C_j(t)\in\mathcal{T}\}\cap\{C_j(0)=\boldsymbol{Client}\}}$$

Thus on taking expectations and dividing by $\mathbb{P}\{C_j(0) = \boldsymbol{Client}\} = \frac{1}{n}\mathbb{E}[N'_{\boldsymbol{C}}(0)]$, we may obtain:

$$\mathbb{P}\{C_j(t) \in \mathcal{T} \mid C_j(0) = \boldsymbol{Client}\} = \frac{\sum_{Q\in\mathcal{T}} \left(\mathbb{E}[N'_Q(t)] - \mathbb{E}[N'_Q(0)]\right)}{\mathbb{E}[N'_{\boldsymbol{C}}(0)]}$$

We now aim to replace the expectation terms in the above by the appropriate ODE approximation. We need to specify the initial conditions for the ODEs associated to the modified absorbing model $\boldsymbol{CS}'(n, m)$ at time 0 in terms of the long-time limits of the ODEs from the original model $\boldsymbol{CS}(n, m)$. This needs to happen according to the preceding discussion regarding how the fluid component counts of $\boldsymbol{CS}'(n, m)$ are initialised from those of $\boldsymbol{CS}(n, m)$ at time 0:

$$(v'_{\boldsymbol{C}}(0), v'_{\boldsymbol{Cw}}(0), v'_{\boldsymbol{Ct}}(0), v'_{\boldsymbol{C}'}(0), v'_{\boldsymbol{Cw}'}(0), v'_{\boldsymbol{Ct}'}(0), v'_{\boldsymbol{S}}(0), v'_{\boldsymbol{Sg}}(0), v'_{\boldsymbol{Sb}}(0)) = (v_{\boldsymbol{C}}, 0, 0, 0, v_{\boldsymbol{Cw}}, v_{\boldsymbol{Ct}}, v_{\boldsymbol{S}}, v_{\boldsymbol{Sg}}, v_{\boldsymbol{Sb}})$$

Then the fluid approximation for the CDF is given by:

$$\mathbb{P}\{C_j(t) \in \mathcal{T} \mid C_j(0) = \boldsymbol{Client}\} \approx \frac{\sum_{Q\in\mathcal{T}} \left(v'_Q(t) - v'_Q(0)\right)}{v_{\boldsymbol{C}}}$$

[4] The system of ODEs usually has infinitely many fixed points, but often has only one that is meaningful in the context of the original model, that is, for example, where the total component population at the fixed point is correct. This is discussed in more detail in Section 4 where conditions guaranteeing convergence of the stationary distribution to the ODE's fixed point for increasing component population size are given.

[5] Where $\mathbf{1}_A$ is the indicator function of the event A.

Fig. 5(b) shows cumulative distribution functions computed using traditional methods for the passage-time random variable discussed above. The cumulative distribution functions converge to the ODE approximation as the component populations increase. This result is proven in general in Section 4.3.2.

General steady-state individual passage times. We will now give a general treatment of steady-state individual passage times by ODE techniques. Furthermore, in order to communicate the main ideas, we omitted some technical detail from the derivation of the example steady-state individual passage-time computation just presented. In particular, the quantity of Eq. (3.9) does not match exactly the standard passage-time random variable definition at steady state. The exact quantity we are in fact interested in is:

$$\mathbb{P}\{C_j(t) \in \mathcal{T} \mid C_j(0) = \boldsymbol{Client}, C_j(0-) = \boldsymbol{Client\_think}\} \tag{3.10}$$

That is to say that time 0 should be the *instant* at which the fluid component under observation performs the *think*-action. Eq. (3.9) does not capture this exactly — rather here we are simply guaranteed that at time 0 the observed *fluid component* has made no further transitions since originally entering the state ***Client***, however, *other components in the model* may well have done. This means that an individual passage-time measurement, which can begin in different model configurations, should actually be specified by target states and source *transitions*, immediately after which the timing starts, rather than source *states*.

In fact for this example, in the limit of large component populations, approximating Eq. (3.10) by Eq. (3.9) is asymptotically correct. Informally, this is because the extra condition of Eq. (3.10) will have no local effect on the fluid component's state, that is, this passage-time example has only one *local* source state ***Client***. Therefore the only effect of the extra conditioning will be on the state of the fluid component's previous cooperation partners. Then, if the maximum number of fluid components which synchronise together in any cooperation stays constant whilst the number of fluid components is scaled up (as is the case for the GPEPA models considered in this paper) it is reasonable that this error will vanish in the fluid limit. The convergence result of Section 4.3.2 will guarantee this formally. In the general case where there can be more than one local source state for the passage time, we will see that extra care is required to ensure an accurate ODE approximation. In particular, we will have to deviate slightly from the approach adopted for the example of the previous section. Specifically, we will consider explicitly ODEs which capture the probability distribution of the observed fluid component directly, whereas, before, these quantities were obtained indirectly by suitable normalisation of the total component-count expectations.

In order to give the general definition of a steady-state individual passage time succinctly, it helps to employ the idea of *memory components*. These are fluid components that can be composed passively with the individual under observation to keep track of its evolution and provide an elegant means of defining the start and end of the passage. They are not allowed to modify the behaviour of the observed component, that is, they may not block it from performing activities. Each steady-state individual passage-time measurement is specified by two such components: one that defines *source transitions* and one that defines *target states*. For example, in the context of the model $\boldsymbol{CS}(n, m)$, the memory fluid component $\boldsymbol{Mem_S}$:

$$\boldsymbol{Mem_S} \stackrel{def}{=} (timeout, \top).\boldsymbol{Mem_S} + (request, \top).\boldsymbol{Mem_{S1}}$$
$$\boldsymbol{Mem_{S1}} \stackrel{def}{=} (timeout, \top).\boldsymbol{Mem_S} + (request, \top).\boldsymbol{Mem_{S1}}$$

can be composed with a ***Client*** component to give the composed fluid component $\boldsymbol{Client} \underset{L_S}{\bowtie} \boldsymbol{Mem_S}$, where $L_S := \{request, timeout\}$. Then the 3-tuple $(\boldsymbol{Mem_{S1}}, \boldsymbol{Mem_S}, timeout)$ specifies that a passage-time measurement starts whenever the memory component makes a transition $\boldsymbol{Mem_{S1}} \xrightarrow{(timeout, \cdot)} \boldsymbol{Mem_S}$ which corresponds to the ***Client*** component having completed a *request*-action followed by a *timeout*-action. Then when a passage-time measurement begins, we consider $\boldsymbol{Client} \underset{L_S}{\bowtie} \boldsymbol{Mem_S}$ composed with a further memory component for which a designated set of target states is given. For example, the memory component:

$$\boldsymbol{Mem_T} \stackrel{def}{=} (think, \top).\boldsymbol{Mem_{T1}}$$
$$\boldsymbol{Mem_{T1}} \stackrel{def}{=} (think, \top).\boldsymbol{Mem_{T1}}$$

together with the designated target state $\boldsymbol{Mem_{T1}}$ asserts that a passage will complete when a *think*-action is observed. More specifically, assuming that the observed component was in the state:

$$(\boldsymbol{Client}' \underset{L_S}{\bowtie} \boldsymbol{Mem_S}') \in ds(\boldsymbol{Client} \underset{L_S}{\bowtie} \boldsymbol{Mem_S})$$

at the beginning of a passage, then the passage completes as soon as the composed component:

$$(\boldsymbol{Client}' \underset{L_S}{\bowtie} \boldsymbol{Mem_S}') \underset{L_T}{\bowtie} \boldsymbol{Mem_T}$$

where $L_T := \{think\}$ enters a derivative state in $(- \underset{L_S}{\bowtie} -) \underset{L_T}{\bowtie} \boldsymbol{Mem_{T1}}$, where $-$ is shorthand for an arbitrary derivative state of the corresponding subcomponent. We now give the general definition of a steady-state individual passage time.

Definition 3.3 (*Steady-state Individual Passage Time*)**.** Let G be a GPEPA model and $(H, P) \in \mathcal{B}(G)$ be the fluid component under observation and assume further that all fluid components in group H are initially P components.

Let Mem_S be a memory fluid component composed with P as $P \underset{L_S}{\bowtie} Mem_S$ specifying the start of the passage by the designated set of transitions $\mathcal{S} \subseteq ds(Mem_S) \times ds(Mem_S) \times \mathcal{A}$. Let Mem_T be a memory fluid component composed as $(P' \underset{L_S}{\bowtie} Mem_S{}') \underset{L_T}{\bowtie} Mem_T$ for $P' \underset{L_S}{\bowtie} Mem_S{}' \in ds(P \underset{L_S}{\bowtie} Mem_S)$ specifying the end of the passage by a designated target state $\mathcal{T} \in ds(Mem_T)$.

Now let G^S be the GPEPA model obtained by replacing the first component P in group H with $P \underset{L_S}{\bowtie} Mem_S$. For this passage-time measure to be meaningful, we assume further that the underlying CTMC of the model G^S has a unique stationary distribution.[6]

Similarly to the earlier examples, we consider the model G^S evolving in its stationary regime and then at some arbitrary time, say 0, we attach the Mem_T memory component. Then by timing how long it takes for the memory component Mem_T to reach its target state $\mathcal{T}$, we can compute the passage-time CDF:

$$\mathbb{P}\{C(t) \in (- \underset{L_S}{\bowtie} -) \underset{L_T}{\bowtie} \mathcal{T} \mid \text{transition in } \mathcal{S} \text{ occurred at time } 0\} \tag{3.11}$$

where $C(t)$ is defined to be the stochastic process which tracks the state of the observed fluid component in the group H as the model evolves.

For example, in order to specify the earlier steady-state individual passage time for $\boldsymbol{CS}(n, m)$, we could use the two straightforward memory fluid components $\boldsymbol{Mem_S} \stackrel{def}{=} (think, \top).\boldsymbol{Mem_S}$ and:

$$\boldsymbol{Mem_T} \stackrel{def}{=} (think, \top).\boldsymbol{Mem_{T1}}$$
$$\boldsymbol{Mem_{T1}} \stackrel{def}{=} (think, \top).\boldsymbol{Mem_{T1}}$$

The transition on which to start a timing a passage is specified by $\mathcal{S} := \{(\boldsymbol{Mem_S}, \boldsymbol{Mem_S}, think)\}$ and the target state is $\mathcal{T} := \boldsymbol{Mem_{T1}}$.

We now proceed to show how a general steady-state individual passage time can be analysed using fluid techniques. As in Section 3.2.1, the first step in the procedure is to compute the fluid approximation to the stationary expectations of the fluid component counts of the model G^S using their associated system of ODEs constructed according to Section 2.1. As before, we achieve this by taking the limit of the ODE solutions as $t \to \infty$, either by numerical integration or direct computation of a fixed point. For any $(Y, Q) \in \mathcal{B}(G^S)$, we will write $v^S_{Y,Q}$ for the approximation to $N^S_{Y,Q} := \lim_{t\to\infty} \mathbb{E}[N^S_{Y,Q}(t)]$ obtained in this manner. Since the ODE solutions approximate component-count expectations and there is a single observed fluid component $P \underset{L_S}{\bowtie} Mem_S$ in the model G^S, we see that the quantities corresponding to derivative states of this fluid component, $v^S_{H,Q}$ for $Q \in ds(P \underset{L_S}{\bowtie} Mem_S)$, approximate the steady-state probability that the observed component is in the given state.[7]

However, in order to capture the passage-time quantity of Eq. (3.11) correctly, we need to compute the steady-state distribution of the observed fluid component *given that a transition in $\mathcal{S}$ has just occurred*. Computing probabilities conditioned on transition instants for a subset of *selected transitions* in this manner is an application of *Palm calculus* applied to stationary continuous-time Markov chains (see, for example, the book by Le Boudec [33, Section 7.5] or Serfozo [34, Section 4.16]). In particular, the probability that the observed component is in state $Q \in ds(P \underset{L_S}{\bowtie} Mem_S)$ immediately after a transition in $\mathcal{S}$ occurs, is the expected rate of enabled transitions in $\mathcal{S}$ which also result in the observed component entering the state Q, divided by the total expected rate of all enabled transitions in $\mathcal{S}$ [e.g. 34, Definition 96 and Proposition 97], that is:

$$\mathbb{P}\{C(0) = Q \mid \text{transition in } \mathcal{S} \text{ occurred at time } 0\}$$
$$= \frac{\sum_{\alpha\in\mathcal{A}} \sum_{C\in ds(P \underset{L_S}{\bowtie} Mem_S)} \left(\frac{1}{r_\alpha(C)} \mathbf{1}_{\mathcal{S}(C,Q,\alpha)} \sum_{C \xrightarrow{(\alpha,\lambda)} Q} \lambda \times \mathbb{E}\left[\mathcal{R}_\alpha(G^S, N^S_0, H, C)\right] \right)}{\sum_{\alpha\in\mathcal{A}} \sum_{C,C'\in ds(P \underset{L_S}{\bowtie} Mem_S)} \left(\frac{1}{r_\alpha(C)} \mathbf{1}_{\mathcal{S}(C,C',\alpha)} \sum_{C \xrightarrow{(\alpha,\lambda)} C'} \lambda \times \mathbb{E}\left[\mathcal{R}_\alpha(G^S, N^S_0, H, C)\right] \right)}$$

where, $N^S_0 \in \mathcal{B}(G^S) \to \mathbb{Z}_+$ represents the state of the underlying CTMC of the model G^S in the stationary regime at time 0 and for $P' \underset{L_S}{\bowtie} Mem_S{}' \in ds(P \underset{L_S}{\bowtie} Mem_S)$ and $P'' \underset{L_S}{\bowtie} Mem_S{}'' \in ds(P \underset{L_S}{\bowtie} Mem_S)$, $\mathbf{1}_{\mathcal{S}(P' \underset{L_S}{\bowtie} Mem_S{}', P'' \underset{L_S}{\bowtie} Mem_S{}'', \alpha)} := 1$ if $(Mem_S{}', Mem_S{}'', \alpha) \in \mathcal{S}$ and is 0 otherwise. The ODE approximation to this expression is obtained by substituting the ODE approximations to the stationary component counts obtained above in place of N^S_0. Specifically, defining $V^S \in \mathcal{B}(G^S) \to \mathbb{Z}_+$

6 Note that this could be implied by irreducibility of the CTMC underlying G^S or, alternatively, if it is reducible but has a single communicating class.

7 It may initially seem surprising that this will provide a good approximation since there is only one copy of the fluid component $P \underset{L_S}{\bowtie} Mem_S$ in the model. However, the key point which ensures that the approximation is good is that this component exists within a large population of identically-distributed P-components. This argument will be given formally in Section 4.3.2.

by $V^S(Y, Q) := v^S_{Y,Q}$, we obtain the ODE approximation:

$$\mathbb{P}\{C(0) = Q \mid \text{transition in } \mathcal{S} \text{ occurred at time } 0\} \approx \frac{\sum_{\alpha\in\mathcal{A}} \sum_{C\in ds(P \underset{L_S}{\bowtie} Mem_S)} \left(\frac{1}{r_\alpha(C)} \mathbf{1}_{\mathcal{S}(C,Q,\alpha)} \sum_{C \xrightarrow{(\alpha,\lambda)} Q} \lambda \times \mathcal{R}_\alpha(G^S, V^S, H, C) \right)}{\sum_{\alpha\in\mathcal{A}} \sum_{C,C'\in ds(P \underset{L_S}{\bowtie} Mem_S)} \left(\frac{1}{r_\alpha(C)} \mathbf{1}_{\mathcal{S}(C,C',\alpha)} \sum_{C \xrightarrow{(\alpha,\lambda)} C'} \lambda \times \mathcal{R}_\alpha(G^S, V^S, H, C) \right)} \tag{3.12}$$

Finally, in order to compute the passage time of interest, we must solve the ODEs associated with the GPEPA model obtained by attaching the memory component Mem_T to the observed fluid component, resulting in fluid components of the form $Q \underset{L_T}{\bowtie} Mem_T$ for $Q \in ds(P \underset{L_S}{\bowtie} Mem_S)$. We will write $v^T_{Y,R}(t)$ for the solutions to this system of ODEs. In order to capture correctly the passage-time quantity starting from a transition instant, the initial condition $v^T_{H,Q \underset{L_T}{\bowtie} Mem_T}(0)$ for an observed component derivative state where $Q \in ds(P \underset{L_S}{\bowtie} Mem_S)$ must be the conditioned quantity computed above in Eq. (3.12). For other observed component derivative states $Q \underset{L_T}{\bowtie} Mem_T{}'$, where $Mem_T{}' \in ds(Mem_T)$ and $Mem_T{}' \neq Mem_T$, the initial condition is 0 since these states track the progress through the passage which has only just started. For all other $(Y, R) \in \mathcal{B}(G^S)$, the initial value $v^T_{Y,R}(0)$ is set to $v^S_{Y,R}$. Then the passage-time CDF is approximated as:

$$\mathbb{P}\{C(t) \in (- \underset{L_S}{\bowtie} -) \underset{L_T}{\bowtie} \mathcal{T} \mid \text{transition in } \mathcal{S} \text{ occurred at time } 0\} \approx \sum_{C\in(- \underset{L_S}{\bowtie} -) \underset{L_T}{\bowtie} \mathcal{T}} v^T_{H,C}(t)$$

4. 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 GPEPA models for increasing total component population. Directly below, we will first define exactly what we mean by a sequence of models with *increasing total component population* by defining the notion of a *structurally-equivalent* sequence of models.

When two GPEPA 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 4.1 (*Structural Equivalence*)**.** Let G_1 and G_2 be two GPEPA 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 $\mathcal{W}(G_1, G_2) = \textbf{true}$, where $\mathcal{W}(\cdot,\cdot)$ is defined by:

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

and **false** in all other cases. Secondly, they must have the same initial fluid component population ratios. That is the initial number of each fluid component derivative state in each component group divided by the model size must be the same for both G_1 and G_2.

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

For a given sequence of structurally-equivalent GPEPA models $\{G^{(i)}\}_{i=1}^{\infty}$ we will sometimes write, for example, $\mathcal{N} := \mathcal{N}(G^{(i)})$, $\mathcal{B} := \mathcal{B}(G^{(i)})$ and $\mathcal{A} := \mathcal{A}(G^{(i)})$ for any i, since these quantities are always well defined by structural equivalence. Write also $\mathbf{N}^{(i)}(t)$ for the aggregated CTMC of $G^{(i)}$ and $\mathbf{v}^{(i)}(t)$ for its differential equation approximation. For the convergence results, we will be concerned primarily with the rescaled quantities $\bar{\mathbf{v}}^{(i)}(t) := \mathbf{v}^{(i)}(t)/\mathcal{S}(G^{(i)})$ and $\bar{\mathbf{N}}^{(i)}(t) := \mathbf{N}^{(i)}(t)/\mathcal{S}(G^{(i)})$. By structural equivalence, the differential equation approximation for each i satisfies the same ODE $\dot{\mathbf{v}}^{(i)}(t) = \mathbf{f}(\mathbf{v}^{(i)}(t))$. Furthermore, by homogeneity (Lemma 2.5), the rescaled quantities $\bar{\mathbf{v}}^{(i)}(t) := \mathbf{v}^{(i)}(t)/\mathcal{S}(G^{(i)})$ also satisfy the same differential equation and are thus independent of i, so we may write just $\bar{\mathbf{v}}(t)$ without ambiguity.

When we wish to specify explicitly a deterministic initial condition for $\mathbf{N}^{(i)}(t)$ we will write $\mathbf{N}^{(i)}_{\mathtt{n}}(t)$ for the process started in state $\mathtt{n} \in \mathbb{Z}_+^{\mathcal{N}}$ corresponding to some aggregate state reachable from the initial state $G^{(i)}$. Similarly, we will also write $\bar{\mathbf{N}}_{\bar{\mathtt{n}}}(t)$ for the rescaled process started in the (rescaled) state $\bar{\mathtt{n}} := \mathtt{n}/\mathcal{S}(G^{(i)})$. We will also use the same subscript notation on ODE solutions to indicate their initial conditions, for example $\bar{\mathbf{v}}_{\bar{\mathtt{n}}}(t)$ for the solution to the initial value problem specified by $\bar{\mathbf{v}}_{\bar{\mathtt{n}}}(0) = \bar{\mathtt{n}}$ and $\dot{\bar{\mathbf{v}}}_{\bar{\mathtt{n}}}(t) = \mathbf{f}(\bar{\mathbf{v}}_{\bar{\mathtt{n}}}(t))$.

In order to proceed with proving convergence of the passage-time approximations, we will first give two generic results regarding the limiting convergence of the fluid approximations of component counts. The first is concerned with transient convergence (Theorem 4.2) and the proof is a fairly straightforward application of fairly well-known results. Indeed, the analogous result has already appeared for PEPA in a number of places [22,10,12,35]. As opposed to those references, where the proof was based on results of Kurtz [36], we employ the more general techniques of Darling and Norris [37].

The second result (Theorem 4.3) is concerned with convergence of the approximation in the steady-state regime. It is the main contribution of this section. We use techniques taken from the related area of stochastic approximation algorithms,

specifically, the work of Benaïm [38]. Such techniques have been used by Benaïm and Le Boudec [7] in the discrete-time case to prove similar results for so-called *mean-field* performance models. Our theorem essentially extends this to the continuous-time case. Finally, new results specific to passage times are then derived from the generic results (Theorems 4.4–4.6).

A precondition for the steady-state convergence will be that the associated system of ODEs has an asymptotically-stable fixed point. We present a methodology for automatically verifying this for a large class of model (so-called *split-free* models) based on a suitable reduction of the differential equation followed by the construction of an appropriate Lyapunov function. This exact same approach could be applied directly to performance models specified by continuous stochastic Petri nets [6] or certain kinds of multi-server queueing networks since the synchronisation dynamics are very similar, resulting in a *piecewise affine* system of ODEs in all of these cases.

4.1. Transient convergence of the fluid approximation

The following theorem establishes that, in the limit of large populations, the probability that a GPEPA model's rescaled stochastic process will exceed a given maximum deviation from its fluid approximation tends to zero. It holds transiently, that is, over finite time horizons $[0, T]$. A similar result regarding the transient convergence of variances to their approximating ODEs (Section 2.2) can be proven [18], but we do not give it here for the sake of brevity.

Theorem 4.2. *Let $\{G^{(i)}\}_{i=1}^{\infty}$ be a sequence of structurally-equivalent GPEPA models where $\mathcal{S}(G^{(i)}) \to \infty$ as $i \to \infty$. Assume the setup and notation introduced above.*

Fix $T \geq 0$ and $\epsilon > 0$. Then as $i \to \infty$:

$$\mathbb{P}\left\{ \sup_{t\in[0,T]} \|\bar{\mathbf{N}}^{(i)}_{\bar{\mathbf{n}}}(t) - \bar{\mathbf{v}}_{\bar{\mathbf{n}}}(t)\| > \delta \right\} \longrightarrow 0$$

uniformly for all initial states $\bar{\mathbf{n}}$.

Alternatively, if for given $\bar{\mathbf{n}}$, $\mathbb{P}\left\{\|\bar{\mathbf{N}}^{(i)}(0) - \bar{\mathbf{n}}\| > \delta\right\} \longrightarrow 0$ as $i \to \infty$ for all $\delta > 0$, we have that as $i \to \infty$, $\mathbb{P}\left\{\sup_{t\in[0,T]} \|\bar{\mathbf{N}}^{(i)}(t) - \bar{\mathbf{v}}_{\bar{\mathbf{n}}}(t)\| > \epsilon\right\} \longrightarrow 0$.

Proof. This theorem can be proved using the methodology of Darling and Norris [37]. In the case of the CTMC $\bar{\mathbf{N}}^{(i)}(t)$, it is sufficient to verify that:

- The jump rate of $\bar{\mathbf{N}}^{(i)}(t)$ is bounded above by $k_1\mathcal{S}(G^{(i)})$ for some $k_1 \in \mathbb{R}_+$;
- The second moment of the jump size of $\bar{\mathbf{N}}^{(i)}(t)$ is bounded above by $k_2\mathcal{S}(G^{(i)})^{-2}$ for some $k_2 \in \mathbb{R}_+$;
- For all i and rescaled CTMC states of $G^{(i)}$, say $\bar{\mathbf{n}}$, it holds that:

$$\mathbf{f}(\bar{\mathbf{n}}) = \lim_{s\to 0} \frac{1}{s}\mathbb{E}[\bar{\mathbf{N}}^{(i)}(t+s) - \bar{\mathbf{N}}^{(i)}(t) | \bar{\mathbf{N}}^{(i)}(t) = \bar{\mathbf{n}}]$$

- The function $\mathbf{f}$ is Lipschitz continuous.

The first condition is satisfied since $\mathcal{S}(G^{(i)})\mathcal{Q}^{\max}(G^{(i)})$ bounds the jump rate of $G^{(i)}$, where $\mathcal{Q}^{\max}(G^{(i)}) := \max_{(H,P)\in\mathcal{B}}\{\sum_{\alpha\in\mathcal{A}} r_\alpha(P)\}$ is the *maximal local rate* and is independent of i. The second condition is also satisfied since at most one fluid component in each group evolves at each jump. The third condition is satisfied since this is equivalent to the definition of $\mathbf{f}$ given in Section 2. Lemma B.1 verifies the final condition. □

4.2. Stationary convergence of the fluid approximation

In this section, we wish to give a version of Theorem 4.2 which is valid as $t \to \infty$, that is, in the *steady-state* or *stationary* regime. Specifically, we will show that, under the right conditions and in the large population limit, the rescaled stationary distributions of a structurally-equivalent sequence of GPEPA models will converge to the unique meaningful fixed point of the approximating system of ODEs. Ding explores the same problem for PEPA models in [12]. Specifically, Ding gives a condition on the underlying CTMCs of a sequence of PEPA models with increasing total component population. If it can be verified, this condition is shown to guarantee simultaneously both asymptotic stability of the ODE fixed point and convergence of the rescaled component-count expectations to the ODE solution. Unfortunately this condition is specified directly in terms of the steady-state probabilities of the underlying CTMCs so, as recognised by the author, its verification is not scalable due to the state-space explosion problem. Therefore it appears just as computationally (in)feasible to compute exactly the steady-state probability vector for the specific model of interest. As it stands then, Ding's result does not offer a *useful* guarantee of the validity of the ODE approximation in the steady-state limit.

In order to proceed, the *worst case* device, Grönwall's lemma [e.g. 39, Page 498], which is usually used to prove transient results such as Theorem 4.2 will need to be replaced with something stronger. Specifically, it is necessary to consider asymptotic stability properties of the approximating system of ODEs. In particular, we will require that the system of ODEs has an asymptotically-stable fixed point, that is, we will need to guarantee that a trajectory starting from any initial condition converges to the fixed point in the limit as $t \to \infty$. Empirically, this appears to be the case more often than not but can

be difficult to guarantee formally.[8] For this reason, before we give the actual convergence theorem in Section 4.2.3, the following two sections are devoted to developing an inexpensive method which will allow formal verification of asymptotic stability of the approximating ODEs for a large class of GPEPA model.

4.2.1. Asymptotic stability on the reachable subset and the reduced ODE

In order for a system of ODEs to have a *globally* asymptotically-stable fixed point, it must at least have a *unique* fixed point. For the system $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$, we should never expect this to be the case when considered over all of $\mathbb{R}^{\mathcal{N}}$. To see why, we can consider again the client–server model $\boldsymbol{CS}(n, m)$ and recall that the system of ODEs approximating the component counts is the same for any value of n and m. However, we would not expect the long-term behaviour of the ODE also to be independent of n and m. For example, consider a system with $n = 100$ clients and the difference between the long-term evolution with $m = 50$ servers and $m = 20$ servers. Indeed, a fixed point of the approximating ODEs (Eq. (A.1)) corresponding to the former case is[9] $(25.28, 12.45, 62.27, 8.09, 37.36, 4.55)^T$ and to the latter case is $(70.11, 4.98, 24.91, 3.24, 14.94, 1.82)^T$. In order to handle this, associated to a GPEPA model G, we will define a subset of $\mathbb{R}^{\mathcal{N}(G)}$ which contains the reachable aggregated state space of G and talk instead of asymptotic stability of an ODE fixed point restricted to that subset. That is, we will be interested in verifying only that all trajectories starting in said subset converge to the fixed point. No such condition will be placed on trajectories starting outside of the subset.

To proceed, we decompose the function $\mathbf{f}(\mathbf{x}) = \mathbf{S}\mathbf{r}(\mathbf{x})$ into the product of a matrix $\mathbf{S} \in \mathbb{R}^{|\mathcal{N}(G)|\times|\mathcal{J}(G)|}$ and a vector-valued function $\mathbf{r} : \mathbb{R}_+^{|\mathcal{N}(G)|} \to \mathbb{R}_+^{|\mathcal{J}(G)|}$. Each column of $\mathbf{S}$ corresponds to an entry of $\mathcal{J}(G)$ (Definition 2.7) and thus a possible way in which fluid components may evolve in synchronisation as part of one transition. In order to define $\mathbf{S}$ formally we must fix an ordering on the elements of $\mathcal{J}(G)$, so write its jth element as $(\mathcal{J}_-^j, \mathcal{J}_+^j, \alpha_j)$. Then the ij-entry of $\mathbf{S}$ is defined as 1 if the ith fluid component in $\mathcal{B}(G)$ (under the same ordering as in Section 2.1) is in $\mathcal{J}_+^j \setminus \mathcal{J}_-^j$, -1 if it is in $\mathcal{J}_-^j \setminus \mathcal{J}_+^j$ and 0 otherwise. Therefore each column specifies the change in each fluid component count when the corresponding synchronisation occurs.[10] We refer to the function $\mathbf{r}(\mathbf{x})$ as the *rate vector* and the jth element of this vector is the quantity $\rho_\alpha(\mathcal{J}_-^j, \mathcal{J}_+^j)\mathcal{R}_\alpha(G, \mathbf{x}, \mathcal{J}_-^j)$, which is the rate at which such a transition occurs.

Now for some GPEPA model G, write $\bar{\mathbf{n}}_0 \in \mathbb{R}^{|\mathcal{N}(G)|}$ for the rescaled vector of component counts corresponding to the model's initial state. Then by the definition of $\mathbf{S}$, we see that any reachable state in the rescaled aggregated CTMC of G is contained in the set $\mathcal{R}(G) := \left\{\bar{\mathbf{n}}_0 + \mathbf{S}\mathbf{x} \,:\, \mathbf{x} \in \mathbb{R}^{|\mathcal{J}(G)|}\right\}$.[11] In the case of a structurally-equivalent sequence of GPEPA models $\{G^{(i)}\}_{i=1}^{\infty}$, $\mathcal{R} := \mathcal{R}(G^{(i)})$ is independent of i and contains the rescaled aggregated state space of all of the models in the sequence. Furthermore for $\bar{\mathbf{r}} = \bar{\mathbf{n}}_0 + \mathbf{S}\mathbf{x} \in \mathcal{R}$, we have:

$$\bar{\mathbf{v}}_{\bar{\mathbf{r}}}(t) = \bar{\mathbf{n}}_0 + \mathbf{S}\mathbf{x} + \int_0^t \mathbf{S}\mathbf{r}(\bar{\mathbf{v}}_{\bar{\mathbf{r}}}(s))\,\mathrm{d}s = \bar{\mathbf{n}}_0 + \mathbf{S}\left(\mathbf{x} + \int_0^t \mathbf{r}(\bar{\mathbf{v}}_{\bar{\mathbf{r}}}(s))\,\mathrm{d}s\right)$$

so that $\bar{\mathbf{v}}_{\bar{\mathbf{r}}}(t) \in \mathcal{R}$ for all $t \in \mathbb{R}_+$. In many cases, the ODEs associated to a structurally-equivalent sequence will have a unique fixed point within $\mathcal{R}$. In the case of the client–server model with ODEs given by Eq. (A.1) and assuming we are interested in the structurally-equivalent sequence $\{\boldsymbol{CS}(2n, n)\}_{n=1}^{\infty}$ (so $\bar{\mathbf{n}}_0 = (2/3, 0, 0, 1/3, 0, 0)^T$), this single fixed point is $\bar{\mathbf{r}}^* = (0.1685, 0.0830, 0.4151, 0.0539, 0.2491, 0.0303)^T$.

In order to apply the convergence theorem of Section 4.2.3, we will need to show that such a fixed point is asymptotically stable within $\mathcal{R}$. That is, that any ODE trajectory starting within $\mathcal{R}$ converges towards the fixed point as $t \to \infty$, or formally, if $\bar{\mathbf{r}} \in \mathcal{R}$, $\lim_{t\to\infty} \bar{\mathbf{v}}_{\bar{\mathbf{r}}}(t) = \bar{\mathbf{r}}^*$. Such a question of asymptotic stability on a subset can be recast as a question of global asymptotic stability on all of $\mathbb{R}^m$ for some m by a suitable reduction of the system of ODEs. Specifically, construct a matrix $\mathbf{R} \in \mathbb{R}^{\mathcal{N}\times m}$ where m is the rank of $\mathbf{S}$ by choosing a maximal set of linearly-independent columns of $\mathbf{S}$. Then let $\mathbf{B} \in \mathbb{R}^{m\times\mathcal{N}}$ be its *Moore–Penrose pseudoinverse* which is defined and unique for all real matrices [43].[12] Since $\mathbf{R}$ has full column rank it then holds that $\mathbf{B}\mathbf{R} = \mathbf{I}_m$ where $\mathbf{I}_m$ is the $m \times m$ identity matrix and, furthermore, the matrix $\mathbf{R}\mathbf{B}$ is an *orthogonal projector* onto the range of $\mathbf{R}$ [43]. In particular it therefore holds that $\mathbf{R}\mathbf{B}\mathbf{S} = \mathbf{S}$.

The system of rescaled *reduced component-counting ODEs* corresponding to a GPEPA model (or structurally-equivalent sequence) is then defined as $\dot{\hat{\mathbf{v}}}(t) = \mathbf{B}\mathbf{S}\mathbf{r}(\mathbf{R}\hat{\mathbf{v}}(t) + \bar{\mathbf{n}}_0)$. Now note that $\mathcal{R} = \left\{\bar{\mathbf{n}}_0 + \mathbf{S}\mathbf{x} \,:\, \mathbf{x} \in \mathbb{R}^{|\mathcal{J}(G)|}\right\} = \{\bar{\mathbf{n}}_0 + \mathbf{R}\mathbf{x} \,:\, \mathbf{x} \in \mathbb{R}^m\}$ and let $\bar{\mathbf{r}} = \bar{\mathbf{n}}_0 + \mathbf{R}\mathbf{x} \in \mathcal{R}$.[13] Then write $\bar{\mathbf{z}}(t) := \mathbf{R}\hat{\mathbf{v}}_{\mathbf{x}}(t) + \bar{\mathbf{n}}_0$. Then it follows that: $\dot{\bar{\mathbf{z}}}(t) = \mathbf{R}\dot{\hat{\mathbf{v}}}_{\mathbf{x}}(t) = \mathbf{R}\mathbf{B}\mathbf{S}\mathbf{r}(\bar{\mathbf{z}}(t)) = \mathbf{S}\mathbf{r}(\bar{\mathbf{z}}(t)$ and furthermore we have that $\bar{\mathbf{z}}(0) = \bar{\mathbf{n}}_0 + \mathbf{R}\mathbf{x}$, so then $\bar{\mathbf{z}}(t) = \bar{\mathbf{v}}_{\bar{\mathbf{r}}}(t) = \mathbf{R}\hat{\mathbf{v}}_{\mathbf{x}}(t) + \bar{\mathbf{n}}_0$ for all $t \in \mathbb{R}_+$. Finally, we have that

8 It is certainly not always the case for GPEPA models; see the counter example in [40].

9 With rate parameter values: $r_r = 2.0, r_{tmt} = 0.3, r_t = 0.2, r_b = 0.1, r_d = 1.0, r_{rst} = 1.0$.

10 In models of chemical reactions, the analogous matrix is often called the *stoichiometric matrix* [41] and has previously been referred to as the *activity matrix* by [3,12] in the context of PEPA.

11 We note that this definition is similar to that of the *linearised reachability set* often used in the context of stochastic Petri nets [42] to encode the reachable discrete state space efficiently for use in various structural verification tasks.

12 Since $\mathbf{R}$ has full rank by construction, its Moore–Penrose pseudoinverse can be computed directly as $(\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T$ if $m < \mathcal{N}$ and is simply $\mathbf{R}^{-1}$ if $m = \mathcal{N}$ [43].

13 Also note that every element of $\mathcal{R}$ has a *unique* decomposition of the form $\bar{\mathbf{n}}_0 + \mathbf{R}\mathbf{x}$ because $\mathbf{R}$ has full column rank.

$\hat{\mathbf{v}}_{\mathbf{x}}(t) = \mathbf{B}(\bar{\mathbf{v}}_{\bar{\mathbf{r}}}(t) - \bar{\mathbf{n}}_0)$. Then $\mathbf{x}^* \in \mathbb{R}^m$ is the unique fixed point of the reduced ODE if and only if $\bar{\mathbf{r}}^* := \bar{\mathbf{n}}_0 + \mathbf{R}\mathbf{x}^*$ is the unique fixed point in $\mathcal{R}$ of the original ODE. Furthermore, $\mathbf{x}^*$ is globally asymptotically stable (over all of $\mathbb{R}^m$) for the reduced ODE if and only if $\bar{\mathbf{r}}^*$ is asymptotically stable restricted to $\mathcal{R}$ for the original ODE.

In the case of the structurally-equivalent sequence $\boldsymbol{CS}(2n, n)$, $\mathbf{S}$ is 6×7 and has rank 4 (the dimension of the reduced system) and the reduced ODE has a unique fixed point[14] $(0.169, 0.083, 0.054, 0.249)^T$.

In the next section we present a technique for verifying global asymptotic stability of the reduced ODE in an inexpensive fashion for a large class of GPEPA models. By the above discussion, this will in turn prove the required asymptotic stability of the original ODE when restricted to the set $\mathcal{R}$.

4.2.2. Verification of global asymptotic stability of the reduced ODE

As discussed, we seek in this section to present an inexpensive technique for verifying global asymptotic stability of the reduced ODE for some GPEPA model G (or structurally-equivalent sequence). By the results of the previous section, this will in turn imply asymptotic stability for the original system of component-counting ODEs restricted to the set $\mathcal{R}(G)$.

A common approach is to construct an appropriate Lyapunov function witnessing the global asymptotic stability [e.g. 44]. For general non-linear systems, this is a very difficult problem, however for the large class of *split-free* GPEPA models,[15] the function $\mathbf{f}$ can be shown to be piecewise linear [2] where the different domains of linearity are separated by hyperplanes. This means that the right-hand side of the reduced system of ODEs $\dot{\hat{\mathbf{v}}}(t) = \mathbf{A}\mathbf{f}(\mathbf{D}\hat{\mathbf{v}}(t) + \mathbf{b})$ is piecewise affine where the regions of different dynamics are also separated by hyperplanes. Fortunately, for piecewise affine systems, there are some inexpensive methods which allow the automatic computation of a Lyapunov function.

One approach is to construct a *common quadratic Lyapunov function* [45] by finding a solution to a set of *linear matrix inequalities* (LMIs). Determining feasible solutions to a system of LMIs is a convex optimisation problem [46], and thus, can be done very efficiently. Specifically, since it has a piecewise-affine right-hand side, the reduced ODE can be written as:

$$\dot{\hat{\mathbf{v}}}(t) = \mathbf{F}_i\hat{\mathbf{v}}(t) + \mathbf{f}_i \quad \text{when } \hat{\mathbf{v}}(t) \in \Gamma_i \text{ for } i = 1, \ldots, f$$

where $\mathbf{F}_i \in \mathbb{R}^{m\times m}$ and $\mathbf{f}_i \in \mathbb{R}^m$ and there are f regions Γ_i separated by hyperplanes. Since the right-hand side is continuous, each such region Γ_i can be chosen to also include the segments of the affine hyperplanes which define its boundaries — we assume here that this is the case.

Assuming the single fixed point of this ODE is $\hat{\mathrm{v}}^* \in \mathbb{R}^m$, we construct the quadratic form $V(\mathbf{y}) := (\mathbf{y} - \hat{\mathrm{v}}^*)^T\mathbf{P}(\mathbf{y} - \hat{\mathrm{v}}^*)$, where $\mathbf{P} \in \mathbb{R}^{m\times m}$ is a symmetric, positive definite matrix solving[16] simultaneously the system of LMIs $\mathbf{P}\mathbf{F}_i + \mathbf{F}_i^T\mathbf{P} + 2\alpha\mathbf{P} < 0$ for $i = 1, \ldots, f$ and some $\alpha > 0$. Where they exist, feasible solutions can be found very quickly using, for example, the MATLAB® LMI toolbox [47]. It can then be shown using the techniques of [45] (see [40] for the full proof details) that the function V is a Lyapunov function which verifies the global asymptotic stability of $\hat{\mathrm{v}}^*$.

It should be noted that the existence of such a common quadratic Lyapunov function is only a sufficient condition for global asymptotic stability. In practice, we have found it to be a very powerful technique which we have not observed to fail. However, there are possible extensions such as *piecewise quadratic Lyapunov functions* [48] which we do not consider in this paper.

Considering again the structurally-equivalent sequence $\boldsymbol{CS}(2n, n)$, the ODE (reduced as in the previous section) gives rise to a system of 4 LMIs. These can be shown feasible in under a second using the MATLAB® toolkit described above on a standard Intel Core 2 Duo Linux machine, thus verifying that their unique fixed point is indeed globally asymptotically stable. As discussed in Section 4.2.1, this in turn verifies that the unique fixed point of the original system of ODEs is also asymptotically stable when restricted to $\mathcal{R}(\boldsymbol{CS}(2n, n))$.

4.2.3. Stationary convergence theorem

Before presenting the convergence theorem we require a few dynamical systems definitions which we give here for the general differential equation $\dot{\mathbf{y}}(t) = \mathbf{k}(\mathbf{y}(t))$ where $\mathbf{y}(t) \in \mathbb{R}^n$ for $t \in \mathbb{R}_+$ and $\mathbf{k} : \mathbb{R}^n \to \mathbb{R}^n$. For more details see any standard text on dynamical systems [e.g. 49].

Let $K \subseteq \mathbb{R}^n$ be compact, then the *omega limit set* of $\mathbf{w} \in K$ with respect to $\mathbf{y}(t)$, denoted $\omega(\mathbf{w}, \mathbf{y}(t))$, is the set of $\mathbf{z} \in \mathbb{R}^n$, such that $\lim_{k\to\infty} \mathbf{y}_{\mathbf{w}}(t_k) = \mathbf{z}$ for some sequence $\{t_k\}_{k=1}^{\infty}$ with $\lim_{k\to\infty} t_k = \infty$. The *Birkhoff centre* of $\mathbf{y}(t)$ in K, denoted $\mathcal{B}(\mathbf{y}(t), K)$, is the closure of the set of *recurrent points* in K, that is, the closure of the set of $\mathbf{w} \in K$ with $\mathbf{w} \in \omega(\mathbf{w}, \mathbf{y}(t))$. It is clear that this set contains any equilibrium points and periodic orbits reachable from K and in the case of an asymptotically stable (within K) fixed point $\mathbf{y}^* \in K$, $\mathcal{B}(\mathbf{y}(t), K) = \{\mathbf{y}^*\}$.

Now since K is compact it is a separable metric space under the restriction of the usual topology on $\mathbb{R}^n$. Assume that $\mathbf{w} \in K$ implies that $\mathbf{y}_{\mathbf{w}}(t) \in K$ holds for all $t \in \mathbb{R}_+$. An *invariant measure* for $\mathbf{y}(t)$ is a probability measure μ on the measurable space

[14] Using the same rates as before and assuming that we have chosen linearly independent columns so as to eliminate the ***Client_think*** and ***Server_broken*** components and that the others are given in the order of their definition as fluid components.

[15] A split-free GPEPA model is one where no synchronised action is enabled by more than one fluid component type on one side of a cooperation between component groups.

[16] Inequalities for matrices, e.g. $Q < 0$ or $Q > 0$ are interpreted as statements of negative or positive definiteness, respectively.

$(K, \mathfrak{B}(K))$[17] such that $\mu(\mathbf{y}_A^{-1}(t)) = \mu(A)$ for every $t \in \mathbb{R}_+$ and $A \in \mathfrak{B}(K)$, where $\mathbf{y}_A^{-1}(t) := \{\mathbf{w} \in K : \mathbf{y}_{\mathbf{w}}(t) \in A\}$. Fix $t > 0$. The *Poincaré recurrence theorem* then says that for any $A \in \mathfrak{B}(K)$:

$$\mu(\{\mathbf{w} \in A : \exists N \geq 1 \text{ such that } \mathbf{y}_{\mathbf{w}}(nt) \notin A \text{ for all } n \geq N\}) = 0$$

Now since K is separable, there exists a countable basis $\{U_i\}_{i=1}^{\infty}$ for its topology. If for each i with $\mathbf{w} \in U_i$ and for any $N \geq 1$, there exists $n \geq N$ such that $\mathbf{y}_{\mathbf{w}}(nt) \in U_i$ then $\mathbf{w} \in \omega(\mathbf{w}, \mathbf{y}(t))$. Define then:

$$V_i := \{\mathbf{w} \in U_i : \forall N \geq 1, \exists n \geq N \text{ such that } \mathbf{y}_{\mathbf{w}}(nt) \in U_i\}$$

and then we have $\mu(V_i) = \mu(U_i)$. Let also $K_i := V_i \cup (K \setminus U_i)$ and then $\mu(K_i) = \mu(K) = 1$. Finally $\cap_{i=1}^{\infty} K_i \subseteq \mathcal{B}(\mathbf{y}(t), K)$ so $\mu(\mathcal{B}(\mathbf{y}(t), K)) = 1$.

The steady-state analogue of Theorem 4.2 now follows. The idea of the proof is to show that any limit point of the sequence of stationary measures of the GPEPA models is invariant with respect to the approximating ODE on a suitable compact set containing the ODE solutions and CTMC state spaces. We will then apply Poincaré's recurrence theorem as above.

Theorem 4.3. *Let $\{G^{(i)}\}_{i=1}^{\infty}$ be a sequence of structurally-equivalent GPEPA models where $\mathcal{S}(G^{(i)}) \to \infty$ as $i \to \infty$. Again, assume the notation introduced above.*

Define the compact set $K := \mathcal{R} \cap \left\{\mathbf{x} : \mathbf{x} \in \mathbb{R}^{\mathcal{N}}, x_i \geq 0, \sum_i x_i \leq 1\right\}$, where $\mathcal{R}$ is as defined in Section 4.2.1. Further assume μ_i is the unique stationary measure of the rescaled CTMC of $G^{(i)}$ which we consider as a probability measure on $(K, \mathfrak{B}(K))$.[18]

Then if $\mathcal{B}(\bar{\mathbf{v}}(t), K) = \{\bar{\mathbf{v}}^*\}$, the sequence of measures μ_n converges in probability and thus also weakly in distribution to the point mass at $\bar{\mathbf{v}}^*$.

We note that the condition $\mathcal{B}(\bar{\mathbf{v}}(t), K) = \{\bar{\mathbf{v}}^*\}$ is implied by global asymptotic stability of the reduced ODE and can thus be verified for a given split-free model using the techniques of the previous two sections. For example, we have already verified this condition for the sequence $\{\boldsymbol{CS}(2n, n)\}_{n=1}^{\infty}$ at the end of Section 4.2.2. Theorem 4.3 thus holds for this sequence of models.

Proof of Theorem 4.3. In this proof, we employ techniques taken from the related area of stochastic approximation algorithms [38].

First we note that the sequence of measures μ_i is tight[19] since the space K is compact. Therefore the sequence μ_i is also relatively compact [50, Theorem 6.1], meaning that every subsequence of μ_i, say μ_{i_j} contains a further subsequence, say $\mu_{i_{j_k}}$ which converges weakly. These limit points are not necessarily equal, that is, the sequence μ_i does not necessarily converge weakly itself. However, we will show that any limit of a subsequence of μ_i is at least an invariant measure for $\bar{\mathbf{v}}(t)$.

Let μ be some weak limit point, say $\mu_{i_j} \xrightarrow{w} \mu$. So we wish to show that if $A \in \mathfrak{B}(K)$ and $t \in \mathbb{R}_+$, we have $\mu(\bar{\mathbf{v}}_A^{-1}(t)) = \mu(A)$. To see this, it is sufficient [50, Theorem 1.3], [51, Lemma 1.22] to verify that for any bounded and continuous $g : K \to \mathbb{R}$:

$$\int_K g(\bar{\mathbf{v}}_{\mathbf{x}}(t))\, \mu(\mathrm{d}\mathbf{x}) = \int_K g(\mathbf{x})\, \mu(\mathrm{d}\mathbf{x}) \tag{4.1}$$

Before we proceed, we observe that for any i:

$$\int_K g(\mathbf{x})\, \mu_i(\mathrm{d}\mathbf{x}) = \int_K \mathbb{E}[g(\bar{\mathbf{N}}_{\mathbf{x}}^{(i)}(t))]\, \mu_i(\mathrm{d}\mathbf{x}) \tag{4.2}$$

This is true since the μ_i are stationary measures of the $\bar{\mathbf{N}}^{(i)}(t)$. To show Eq. (4.1), fix $\delta > 0$, then combining Eq. (4.2) with the fact that $\mu_{i_j} \xrightarrow{w} \mu$, we may choose J sufficiently large such that for all $j \geq J$:

$$\left| \int_K g(\mathbf{x})\, \mu(\mathrm{d}\mathbf{x}) - \int_K \mathbb{E}[g(\bar{\mathbf{N}}_{\mathbf{x}}^{(i_j)}(t))]\, \mu_{i_j}(\mathrm{d}\mathbf{x}) \right| < \delta/4 \tag{4.3}$$

We now proceed to bound the term $|\mathbb{E}[g(\bar{\mathbf{N}}_{\mathbf{x}}^{(i)}(t))] - g(\bar{\mathbf{v}}_{\mathbf{x}}(t))|$ for large enough i, uniformly for any rescaled state $\mathbf{x}$ of $G^{(i)}$. We note that g is uniformly continuous since K is compact. Therefore we can find $\alpha > 0$ independent of $\mathbf{x}$ such that $\|\mathbf{y} - \bar{\mathbf{v}}_{\mathbf{x}}(t)\| < \alpha \Rightarrow |g(\mathbf{y}) - g(\bar{\mathbf{v}}_{\mathbf{x}}(t))| < \delta/4$. This allows us to employ the following straightforward upper bound, where $\|g\|$ is an upper bound on the magnitude of g:

$$\left|\mathbb{E}[g(\bar{\mathbf{N}}_{\mathbf{x}}^{(i)}(t))] - g(\bar{\mathbf{v}}_{\mathbf{x}}(t))\right| \leq \delta/4 + 2\|g\|\mathbb{P}\left\{\sup_{0 \leq s \leq t} \|\bar{\mathbf{N}}_{\mathbf{x}}^{(i)}(s) - \bar{\mathbf{v}}_{\mathbf{x}}(s)\| \geq \alpha\right\}$$

17 Where $\mathfrak{B}(K)$ is the Borel σ-algebra of K.

18 This is possible because $\mathcal{R}$ contains the rescaled state space of all of the $G^{(i)}$ and the sum of their rescaled component counts are always bounded by 0 and 1.

19 See for example the book by Billingsley [50, Page 37] for a definition of tightness of measures and also for the general theory of weak convergence of measures.

Then applying Theorem 4.2 , we may choose M sufficiently large such that for all $m \geq M$, $|\mathbb{E}[g(\bar{\mathbf{N}}_{\mathbf{x}}^{(m)}(t))] - g(\bar{\mathbf{v}}_{\mathbf{x}}(t))| \leq \delta/2$ uniformly for any $\mathbf{x}$ in the rescaled state space of $G^{(m)}$. Choose N sufficiently large such that $N \geq J$ and $i_N \geq M$, then, using this bound and that of Eq. (4.3), we obtain for all $n \geq N$:

$$\begin{aligned}\left|\int_K g(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x}) - \int_K g(\bar{\mathbf{v}}_{\mathbf{x}}(t))\,\mu_{i_n}(\mathrm{d}\mathbf{x})\right| &\leq \left|\int_K g(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x}) - \int_K \mathbb{E}[g(\bar{\mathbf{N}}_{\mathbf{x}}^{(i_n)}(t))]\,\mu_{i_n}(\mathrm{d}\mathbf{x})\right| \\ &\quad + \int_K |\mathbb{E}[g(\bar{\mathbf{N}}_{\mathbf{x}}^{(i_n)}(t))] - g(\bar{\mathbf{v}}_{\mathbf{x}}(t))|\,\mu_{i_n}(\mathrm{d}\mathbf{x}) \\ &\leq 3\delta/4 \end{aligned} \tag{4.4}$$

Also since $g(\bar{\mathbf{v}}_{\mathbf{x}}(t))$ is bounded and continuous as a function of $\mathbf{x}$,[20] we can use the fact that $\mu_{ij} \xrightarrow{w} \mu$ to find R sufficiently large such that for all $r \geq R$:

$$\left|\int_K g(\bar{\mathbf{v}}_{\mathbf{x}}(t))\,\mu(\mathrm{d}\mathbf{x}) - \int_K g(\bar{\mathbf{v}}_{\mathbf{x}}(t))\,\mu_{i_r}(\mathrm{d}\mathbf{x})\right| < \delta/4$$

Then combining this with Eq. (4.4), we can obtain:

$$\left|\int_K g(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x}) - \int_K g(\bar{\mathbf{v}}_{\mathbf{x}}(t))\,\mu(\mathrm{d}\mathbf{x})\right| < \delta$$

Since δ was arbitrary, this verifies Eq. (4.1) and thus shows that μ is an invariant measure for $\bar{\mathbf{v}}(t)$.

Now let $C \subset K$ be closed and disjoint from $\mathcal{B}(\bar{\mathbf{v}}(t), K) = \{\bar{v}^*\}$. Then by the *Poincaré recurrence theorem*, $\mu(C) = 0$. This holds for all limit points, μ, of μ_i.

Finally, we assume for a contradiction that $\lim_{i\to\infty} \mu_i(C) \neq 0$. Then there is some $\epsilon > 0$ such that $\mu_{i_l}(C) > \epsilon$ for all elements of some subsequence μ_{i_l}. By relative compactness, we can find a further subsequence, say $\mu_{i_{l_q}}$, such that $\mu_{i_{l_q}} \xrightarrow{w} \mu$ for some weak limit point μ. But since C is closed, we have $\limsup_{q\to\infty} \mu_{i_{l_q}}(C) \leq \mu(C) = 0$, a contradiction, which gives the required result. □

4.3. Convergence of passage-time approximations

In this section, we present the convergence results for global and individual passage times to their differential-equation approximations.

4.3.1. Global passage times

Let $\{G^{(i)}\}_{i=1}^{\infty}$ be a sequence of structurally-equivalent GPEPA models with associated aggregated stochastic processes $\mathbf{N}^{(i)}(t)$ and associated ODE $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$. 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 consider the sequence of global passage times $\{\sigma_i\}_{i=1}^{\infty}$ (Definition 3.1) defined in terms of the rescaled processes by:

$$\sigma_i := \inf\{t \in \mathbb{R}_+ : \mathbf{c} \cdot \bar{\mathbf{N}}^{(i)}(t) \leq C\}$$

with differential-equation approximation $\gamma := \inf\{t \in \mathbb{R}_+ : \mathbf{c} \cdot \bar{\mathbf{v}}(t) \leq C\}$. The following theorem then gives the desired convergence in probability result. See Fig. 2 for an example of this convergence.

Theorem 4.4. *Let $\{G^{(i)}\}_{i=1}^{\infty}$ be a sequence of structurally-equivalent GPEPA models where $\mathcal{S}(G^{(i)}) \to \infty$ as $i \to \infty$. Assume the setup and notation introduced 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. Fix $T > \gamma$ and $\epsilon < \min(\gamma, T - \gamma)$. Choose $\delta^- := \mathbf{c}.\bar{\mathbf{v}}(\gamma - \epsilon) - C > 0$ and $\delta^+ := C - \mathbf{c}.\bar{\mathbf{v}}(\gamma + \epsilon) > 0$. Then:

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

The result then follows by Theorem 4.2. □

4.3.2. Individual passage times

We now turn to proving convergence results for individual passage-times.

[20] See, for example, [24] for the relevant arguments regarding continuous dependence on initial conditions.

Individual passage times starting immediately. Let $\{G^{(i)}\}_{i=1}^{\infty}$ be a sequence of structurally-equivalent GPEPA models with associated aggregated stochastic processes $\mathbf{N}^{(i)}(t)$ and associated ODE $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$. Let $(H, P) \in \mathcal{B}$ be the fluid component under observation and let $\mathcal{T} \subseteq ds(P)$ be the absorbing set of target states specifying the passage-time measurement. Then consider the sequence of transient individual passage times $\{\theta_i\}_{i=1}^{\infty}$ (Definition 3.2) defined by:

$$\theta_i := \inf\{t \in \mathbb{R}_+ \,:\, C^{(i)}(t) \in T\}$$

where $C^{(i)}(t) \in ds(P)$ tracks the state of one of the initial P components in group H of $G^{(i)}$. Recall that:

$$\mathbb{P}\{\theta_i \leq t\} = \mathbb{P}\{C^{(i)}(t) \in \mathcal{T}\} = \frac{1}{\mathcal{S}(G^{(i)}, H)} \sum_{Q \in \mathcal{T}} \mathbb{E}[N^{(i)}_{H,Q}(t)] = k \sum_{Q \in \mathcal{T}} \mathbb{E}[\bar{N}^{(i)}_{H,Q}(t)] \tag{4.5}$$

for $k := \frac{\mathcal{S}(G^{(i)})}{\mathcal{S}(G^{(i)},H)}$ independent of i by structural equivalence. The differential-equation approximation to the CDF can then be expressed independently of i in terms of the rescaled ODE solution:

$$\mathbb{P}\{\theta_i \leq t\} \approx k \sum_{Q \in \mathcal{T}} \bar{v}_{H,Q}(t) \tag{4.6}$$

Theorem 4.5. *Let $\{G^{(i)}\}_{i=1}^{\infty}$ be a sequence of structurally-equivalent GPEPA models where $\mathcal{S}(G^{(i)}) \to \infty$ as $i \to \infty$. Assume the setup and notation introduced above.*

Fix $T > 0$. Then, uniformly for any $t \in [0, T]$, as $i \to \infty$:

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

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

$$\|\mathbb{E}[\bar{\mathbf{N}}^{(i)}(t)] - \bar{\mathbf{v}}(t)\|^2 \leq \mathbb{E}[\|\bar{\mathbf{N}}^{(i)}(t) - \bar{\mathbf{v}}(t)\|^2] \leq \left[\mathbb{P}\left\{\sup_{t \in [0,T]} \|\bar{\mathbf{N}}^{(i)}(t) - \bar{\mathbf{v}}(t)\|^2 \geq \delta\right\} \mathcal{N} + \delta\right]$$

By Theorem 4.2, the limit of this quantity as $i \to \infty$ is δ. The required result follows since this holds for any $\delta > 0$. □

This theorem is illustrated by Fig. 5(a).

Steady-state individual passage times. Let $\{G^{(i)}\}_{i=1}^{\infty}$ be a sequence of structurally-equivalent GPEPA models with associated aggregated stochastic processes $\mathbf{N}^{(i)}(t)$ and associated ODE $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$. Let $(H, P) \in \mathcal{B}$ be the fluid component under observation and assume that all components in group H are initially P components. Further, let Mem_S be the memory fluid component specifying the start of the passage by the designated set of transitions $\mathcal{S} \subseteq ds(Mem_S) \times ds(Mem_S) \times \mathcal{A}$. Let Mem_T be the memory fluid component specifying the end of the passage by the designated target state $\mathcal{T} \in ds(Mem_T)$.

Let $\{G^{S,(i)}\}_{i=1}^{\infty}$ be the sequence of GPEPA models obtained by replacing the first P component in group H of $G^{(i)}$ by $P \underset{L_S}{\bowtie} Mem_S$. In line with Definition 3.3, we assume that the underlying CTMC of each model $G^{S,(i)}$ has a unique stationary distribution and then consider the model $G^{S,(i)}$ evolving in its stationary regime. At some arbitrary time, say 0, we attach the Mem_T memory component. By timing how long it takes for the memory component Mem_T to reach its target state $\mathcal{T}$, we obtain a sequence of passage-time CDFs:

$$\mathbb{P}\{C^{(i)}(t) \in (- \underset{L_S}{\bowtie} -) \underset{L_T}{\bowtie} \mathcal{T} \mid \text{transition in } \mathcal{S} \text{ occurred at time } 0\}$$

where $C^{(i)}(t)$ is defined to be the stochastic process which tracks the state of the observed fluid component in the group H as the model $G^{S,(i)}$ evolves.

Write $\mathcal{B}^S := \mathcal{B}(G^{S,(i)})$ which is independent of i. Before proving the convergence theorem, we first develop the fluid approximation in the context of the sequences of models and show that, similarly to the previous results, the approximation is the same for all models in the sequence. We will find the following observation useful. For $V^{S,(i)} \in \mathcal{B}^S \to \mathbb{R}_+$, the component rate function (Definition 2.1) has the following form for fluid components $R \in ds(P \underset{L_S}{\bowtie} Mem_S)$:

$$\mathcal{R}_\alpha(G^{S,(i)}, V^{S,(i)}, H, R) = \mathcal{R}_{H,R} \times V^{S,(i)}(H, R) \tag{4.7}$$

where $\mathcal{R}_{H,R}$ depends on $V^{S,(i)}$ only through the rescaled component counts:

$$\begin{aligned}
&\frac{1}{\mathcal{S}(G^{(i)})}\left(V^{S,(i)}(H, U) + \sum_{O \in U \underset{L_S}{\bowtie} -} V^{S,(i)}(H, O)\right) \quad \text{for } (H, U) \in \mathcal{B} \\
&\frac{1}{\mathcal{S}(G^{(i)})} V^{S,(i)}(Y, U) \quad \text{for } (Y, U) \in \mathcal{B}, Y \neq H
\end{aligned} \tag{4.8}$$

In order for the fluid approximation to these passage times to be well defined, we assume that the system of ODEs $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$ has a unique rescaled fixed point that is meaningful in the sense that it lies in K (as defined in Theorem 4.3). We represent this rescaled fixed point by $\bar{V} \in \mathcal{B} \to \mathbb{R}_+$.

We will write $\dot{\mathbf{v}}^S(t) = \mathbf{f}^S(\mathbf{v}^S(t))$ for the ODEs associated to any $G^{S,(i)}$ (it is straightforward to see that the function $\mathbf{f}^S$ is the same for all i). A fixed point, say $V^{S,(i)} \in \mathcal{B}^S \to \mathbb{R}_+$, for this system is meaningful in the context of the model $G^{S,(i)}$ if $V^{(i)} \in \mathcal{B} \to \mathbb{R}_+$ defined for $(Y, Q) \in \mathcal{B}$ by:

$$V^{(i)}(Y,Q) := \frac{1}{\mathcal{S}(G^{(i)})} \begin{cases} V^{S,(i)}(H,Q) + \displaystyle\sum_{R \in Q \underset{L_S}{\bowtie} -} V^{S,(i)}(H,R) & : Y = H \\ V^{S,(i)}(Y,Q) & : \text{otherwise} \end{cases}$$

lies in K; and for each $R \in ds(P \underset{L_S}{\bowtie} Mem_S)$, $V^{S,(i)}(H,R) \geq 0$ and $\sum_{R \in ds(P \underset{L_S}{\bowtie} Mem_S)} V^{S,(i)}(H,R) = 1$. We show now how such a fixed point $V^{S,(i)}$ can be constructed from $\bar{V}$. Let $Q \in ds(P \underset{L_S}{\bowtie} Mem_S)$ and note that, in this case, by the above observation regarding the component rate function:

$$f^S_{H,Q}(V^{S,(i)}) = \sum_{R \in ds(P \underset{L_S}{\bowtie} Mem_S)} A_{R,Q} \times V^{S,(i)}(H,R) \tag{4.9}$$

where the $A_{R,Q}$ depend on $V^{S,(i)}$ only through the rescaled component counts given in Eq. (4.8). It thus makes sense to equate the two quantities of Eq. (4.8) with $\bar{V}(H,U)$ and $\bar{V}(Y,U)$, respectively. Then setting Eq. (4.9) to zero for each $Q \in ds(P \underset{L_S}{\bowtie} Mem_S)$ forms a linear system of equations, say $\mathbf{x}^T\mathbf{A} = \mathbf{0}$, where $\mathbf{A} := (A_{R,Q})$ is the $|ds(P \underset{L_S}{\bowtie} Mem_S)| \times |ds(P \underset{L_S}{\bowtie} Mem_S)|$ real matrix of this linear system which is easily seen to have the form of a generator matrix of a CTMC. Indeed, it is the matrix of the 'marginal CTMC' underlying $P \underset{L_S}{\bowtie} Mem_S$ in the context of the model $G^{S,(i)}$ where the rates of actions on which the component cooperates with the rest of the model are given by the values of the relevant components of the fixed point $\bar{V}$. For this reason, this linear system must have a unique solution that is a probability vector or, otherwise, the assumption that the CTMC underlying $G^{S,(i)}$ has a unique stationary distribution would be contradicted. For $Q \in ds(P \underset{L_S}{\bowtie} Mem_S)$, we thus define $V^{S,(i)}(H,Q)$ to be this unique probability vector solution. Together with the above, this fully determines the quantity $V^{S,(i)}$. It is straightforward to verify that this is indeed a meaningful fixed point of the system $\dot{\mathbf{v}}^S(t) = \mathbf{f}^S(\mathbf{v}^S(t))$.

For a given $G^{S,(i)}$, it is also clear that each distinct meaningful fixed point of $\dot{\mathbf{v}}^S(t) = \mathbf{f}^S(\mathbf{v}^S(t))$ yields a distinct meaningful fixed point of $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$ so, for each $G^{S,(i)}$, $\dot{\mathbf{v}}^S(t) = \mathbf{f}^S(\mathbf{v}^S(t))$ must have a unique meaningful fixed point. Furthermore, for $Q \in ds(P \underset{L_S}{\bowtie} Mem_S)$, we observe that the quantities $V^{S,(i)}(H,Q)$ were in fact constructed independently of i and thus may be written as $V^S(H,Q)$.

In order to proceed with the fluid approximation, we need to compute the approximation to the transition instant distributions as given in Eq. (3.12) of Section 3.2.1. Specifically, we see that for $Q \in ds(P \underset{L_S}{\bowtie} Mem_S)$, the approximation is given by:

$$\mathbb{P}\{C^{(i)}(0) = Q \mid \text{transition in } \mathcal{S} \text{ occurred at time } 0\} \approx \frac{\sum_{\alpha \in \mathcal{A}} \sum_{C \in ds(P \underset{L_S}{\bowtie} Mem_S)} \left(\frac{1}{r_\alpha(C)} \mathbf{1}_{\mathcal{S}(C,Q,\alpha)} \sum_{C \xrightarrow{(\alpha,\lambda)} Q} \lambda \times \mathcal{R}_\alpha(G^{S,(i)}, V^{S,(i)}, H, C) \right)}{\sum_{\alpha \in \mathcal{A}} \sum_{C,C' \in ds(P \underset{L_S}{\bowtie} Mem_S)} \left(\frac{1}{r_\alpha(C)} \mathbf{1}_{\mathcal{S}(C,C',\alpha)} \sum_{C \xrightarrow{(\alpha,\lambda)} C'} \lambda \times \mathcal{R}_\alpha(G^{S,(i)}, V^{S,(i)}, H, C) \right)} \tag{4.10}$$

Now by the earlier observation, this expression also depends on $V^{S,(i)}$ only through the rescaled component counts of Eq. (4.8) and the quantities $V^S(H,R)$ for $R \in ds(P \underset{L_S}{\bowtie} Mem_S)$. It is thus independent of i.

Finally, in order to compute the approximate CDF, we recall from Section 3.2.1, that we must solve the ODEs corresponding to the model $G^{S,(i)}$ after the Mem_T memory component has been attached, with initial conditions derived appropriately from $V^{S,(i)}$. Let $V^{T,(i)}_t$ represent the unique solution to these ODEs with the initial condition $V^{T,(i)}_0(H, Q \underset{L_T}{\bowtie} Mem_T)$ for $Q \in ds(P \underset{L_S}{\bowtie} Mem_S)$ given by the quantity of Eq. (4.10), $V^{T,(i)}_0(H, Q \underset{L_T}{\bowtie} Mem_T{}') := 0$ for $Q \in ds(P \underset{L_S}{\bowtie} Mem_S)$ where $Mem_T{}' \in ds(Mem_T)$ and $Mem_T{}' \neq Mem_T$, and for all other $(Y,R) \in \mathcal{B}^S$, the initial value $V^{T,(i)}_0(Y,R) := V^{S,(i)}(Y,R)$. Then we note the solution $\bar{V}_t \in \mathcal{B} \to \mathbb{R}_+$ to the ODEs $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$ given by the initial condition $\bar{V}_0 := \bar{V}$ is such that for all $t \in [0,T]$ and $(Y,Q) \in \mathcal{B} \to \mathbb{R}_+$, in the limit as $i \to \infty$:

$$\bar{V}_t(Y,Q) \longleftarrow \frac{1}{\mathcal{S}(G^{(i)})} \begin{cases} V^{T,(i)}_t(H,Q) + \displaystyle\sum_{R \in (Q \underset{L_S}{\bowtie} -)-} V^{T,(i)}_t(H,R) & : Y = H \\ V^{T,(i)}_t(Y,Q) & : \text{otherwise} \end{cases}$$

and thus since $\bar{V}$ is a fixed point of $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$, we have, for all $t \in \mathbb{R}_+$ and $(Y, Q) \in \mathcal{B}$, as $i \to \infty$:

$$\frac{1}{\mathcal{S}(G^{(i)})}\left(V_t^{T,(i)}(H,Q) + \sum_{R \in (Q \underset{L_S}{\bowtie} -)-} V_t^{T,(i)}(H,R)\right) \longrightarrow \bar{V}(Y,Q) \quad \text{for } Y = H$$

$$\frac{1}{\mathcal{S}(G^{(i)})} V_t^{T,(i)}(Y,Q) \longrightarrow \bar{V}(Y,Q) \quad \text{otherwise} \tag{4.11}$$

Finally, we observe that for $R \in ds(Q \underset{L_T}{\bowtie} Mem_T)$ where $Q \in ds(P \underset{L_S}{\bowtie} Mem_S)$, the corresponding ODE has the form:

$$\dot{V}_t^{T,(i)}(H,R) = \sum_{U \in (- \underset{L_S}{\bowtie} -) \underset{L_T}{\bowtie} -} B_{U,R}^{(i)} \times V_t^{T,(i)}(H,U)$$

where, in the limit $i \to \infty$, the $B_{U,R}^{(i)}$ depends on $V_t^{T,(i)}$ only through the constant rescaled quantities of Eq. (4.11). It thus follows that the quantities $V_t^{T,(i)}(H,R)$ have a limit as $i \to \infty$ for all $t \in [0,T]$, which we write as $V_t^T(H,R)$. It then follows immediately that the CDF approximation $\sum_{C \in (- \underset{L_S}{\bowtie} -) \underset{L_T}{\bowtie} \mathcal{T}} V_t^T(H,C)$ is thus independent of i for all $t \in \mathbb{R}_+$.

Theorem 4.6. *Let $\{G^{(i)}\}_{i=1}^{\infty}$ be a sequence of structurally-equivalent GPEPA models where $\mathcal{S}(G^{(i)}) \to \infty$ as $i \to \infty$. Assume the setup and notation introduced above.*

Fix $T > 0$. Further assume that the system of ODEs $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$ has an asymptotically stable fixed point within K (as defined in Theorem 4.3), that is, $|\mathcal{B}(\mathbf{v}(t), K)| = 1$.

Then we have uniformly for any $t \in [0,T]$, as $i \to \infty$:

$$\mathbb{P}\{C^{(i)}(t) \in (- \underset{L_S}{\bowtie} -) \underset{L_T}{\bowtie} \mathcal{T} \mid \textit{transition in } \mathcal{S} \textit{ occurred at time } 0\} \longrightarrow \sum_{C \in (- \underset{L_S}{\bowtie} -) \underset{L_T}{\bowtie} \mathcal{T}} V_t^T(H,C)$$

The theorem is illustrated by Figs. 5(b), 7(b), 8 and 9.

Recall that for a split-free model, the techniques of Sections 4.2.1 and 4.2.2 can be used to verify the condition of asymptotic stability required by this theorem. We remind the reader that although these techniques deliver only a sufficient condition for asymptotic stability, in practice we have found it to be a very powerful technique. We have not yet found any cases where we suspect asymptotic stability but this approach is not able to verify it. In the next section, we consider a much more detailed case study and use these techniques to verify asymptotic stability in this much more realistic case. Should we require a stronger test in the future, however, possible directions include the construction of *piecewise quadratic Lyapunov functions* [48].

In the case of splitting models, these approaches do not work. It may, in some cases, however, be possible to construct an ad-hoc Lyapunov function to verify asymptotic stability. Otherwise, numerical experiments can be performed to build confidence empirically that a unique meaningful fixed point is asymptotically stable by testing a number of trajectories numerically for convergence to the fixed point.

Proof of Theorem 4.6. First we note that the CTMC underlying each $G^{(i)}$ has a unique stationary distribution since this is true of each $G^{S,(i)}$. Let $\bar{N}^{(i)} \in \mathcal{B} \to \mathbb{R}_+$ be the random variable given by the rescaled component counts of $G^{(i)}$ at time 0. Then by Theorem 4.3, as $i \to \infty$, we have that $\bar{N}^{(i)}$ converges in probability to the ODE fixed point $\bar{V}$. Let $N^{S,(i)} \in \mathcal{B}^S \to \mathbb{R}_+$ be the random variable given by the component counts of $G^{S,(i)}$ at time 0. Also define $\bar{N}^{S,(i)} \in \mathcal{B} \to \mathbb{R}_+$ as follows, for $(Y,Q) \in \mathcal{B}$:

$$\bar{N}^{S,(i)}(Y,Q) := \begin{cases} \dfrac{1}{\mathcal{S}(G^{(i)})}\left(N^{S,(i)}(H,Q) + \displaystyle\sum_{R \in Q \underset{L_S}{\bowtie} -} N^{S,(i)}(H,R)\right) & : Y = H \\ \dfrac{1}{\mathcal{S}(G^{(i)})} N^{S,(i)}(Y,Q) & : \text{otherwise} \end{cases}$$

Then since $\bar{N}^{S,(i)}(Y,Q)$ is equal in distribution to $\bar{N}^{(i)}(Y,Q)$, it is immediate that $\bar{N}^{S,(i)}(Y,Q)$ converges in probability to $\bar{V}(Y,Q)$. We wish now to show that this convergence in probability still holds when conditioned on a transition in $\mathcal{S}$ having just occurred at time 0. For $(Y,R) \in \mathcal{B}$, the mean of $\bar{N}^{S,(i)}(Y,R)$ conditioned on such a transition can be computed as [e.g. 34, Definition 96 and Proposition 97]:

$$\frac{\sum_{\alpha \in \mathcal{A}} \sum_{C \in ds(P \underset{L_S}{\bowtie} Mem_S)} \left(\frac{1}{r_\alpha(C)} \mathbf{1}_{\mathcal{S}(C,Q,\alpha)} \sum_{C \xrightarrow{(\alpha,\lambda)} Q} \lambda \times \mathbb{E}[(\bar{N}^{S,(i)}(Y,R) + S^R_{C,Q}) \mathcal{R}_\alpha(G^{S,(i)}, N^{S,(i)}, H, C)]\right)}{\sum_{\alpha \in \mathcal{A}} \sum_{C,C' \in ds(P \underset{L_S}{\bowtie} Mem_S)} \left(\frac{1}{r_\alpha(C)} \mathbf{1}_{\mathcal{S}(C,C',\alpha)} \sum_{C \xrightarrow{(\alpha,\lambda)} C'} \lambda \times \mathbb{E}[\mathcal{R}_\alpha(G^{S,(i)}, N^{S,(i)}, H, C)]\right)}$$

where $S^R_{C,Q}$ defined as $1/\mathcal{S}(G^{(i)})$ if $Q = R \neq C$, $-1/\mathcal{S}(G^{(i)})$ if $C = R \neq Q$ and 0 otherwise. By recalling the observation of Eq. (4.7), it is straightforward to show that this converges to $\bar{V}(Y, R)$ as $i \to \infty$. A similar approach can be used to show that the variance of $\bar{N}^{S,(i)}(Y, R)$ conditioned on a transition in $\mathcal{S}$ converges to zero as $i \to \infty$. It thus follows that the desired convergence in probability still holds conditioned on such a transition.

Next we wish to show that for $Q \in ds(P \underset{L_S}{\bowtie} Mem_S)$, $\mathbb{E}[N^{S,(i)}(H, Q)]$ converges to $V^S(H, Q)$ as $i \to \infty$. Dynkin's formula [e.g. 52, Page 254] gives:

$$\mathbb{E}[f^S_{H,Q}(N^{S,(i)})] = 0$$

and we recall that due to the form of $\mathbf{f}^S$, we may write this as:

$$\sum_{R \in ds(P \underset{L_S}{\bowtie} Mem_S)} \mathbb{E}[B_{R,Q} \times N^{S,(i)}(H, R)] = 0$$

where the $B_{R,Q}$ depend on $N^{S,(i)}$ only through the quantities $\bar{N}^{S,(i)}$. It thus follows that:

$$\sum_{R \in ds(P \underset{L_S}{\bowtie} Mem_S)} A_{R,Q} \times \mathbb{E}[N^{S,(i)}(H, R)] \longrightarrow 0$$

as $i \to \infty$, where the matrix $\mathbf{A} := (A_{R,Q})$ is as defined earlier in Eq. (4.9). Then since the linear equation $\mathbf{x}^T\mathbf{A} = \mathbf{0}$ was shown above to have a unique solution that is a probability vector, we have that $\mathbb{E}[N^{S,(i)}(H, Q)] \longrightarrow V^S(H, Q)$ as $i \to \infty$. It then follows also that the conditional expectation of $N^{S,(i)}(H, Q)$ given that a transition in $\mathcal{S}$ occurred at time 0 converges to the quantity of Eq. (4.10).

The proof is then concluded by applying Theorem 4.2, Dynkin's formula and Grönwall's lemma [e.g. 39, Page 498] to obtain the final transient convergence result for the CDFs. □

5. Example: customer–service system

We demonstrate the new passage-time analysis techniques on a larger example. We consider a large abstracted customer–service system with two classes of customers, service preemption and failure. The customers from each class require access to one of the services before completing their phase of service. Such systems, with very large numbers of customers and services could for example represent a virtualised service infrastructure, a large parallel architecture running behind internet search engines or massive multimedia content providers.

The provider of the services can offer different *service level agreements* (SLAs) to the customers from each class. Often, these are expressed using the individual passage times defined in Section 3.2. For example, the provider can guarantee that customers in one of the classes (say, ***H*** for *high priority*) will finish their think action within 8 seconds at least 95% of the time and that the customers in the other class (say, ***L*** *low priority*) will finish their think action within 40 s at least 80% of the time. The fluid passage-time analysis techniques described in Section 3.2 are able to verify efficiently whether the two SLAs are satisfied.

As mentioned previously, the key advantage of the fluid-analysis approach is that it can analyse the large system configurations *very rapidly*. This means that large parameter spaces can be explored, even when applying just naïve sweeping methods. For example, a provider of the customer–service system could try to minimise the number of active services while still guaranteeing the SLAs. By applying fluid-analysis techniques to different combinations of scheduling policy parameters, values can be found which both minimise the number of services and satisfy the SLAs, with relatively low computational cost.

We give a GPEPA definition of a moderately complex customer–service system and demonstrate how to apply the techniques from Sections 3.1 and 3.2 to access the global passage times and the individual passage times underlying the above SLAs. We use the Grouped PEPA Analyser (GPA) [18] to show how we can use the rapid computation of the SLAs over many system parametrisations to answer various scalability questions.

The two classes of customers are represented by the fluid components $\boldsymbol{Customer^L}$ and $\boldsymbol{Customer^H}$. Each customer sends a request to the service and waits for a response. Additionally, the high priority customers are allowed to switch to a mode where they negotiate a preemption of a low priority customer:

$$\boldsymbol{Customer^L} \stackrel{def}{=} (request^L, r_{request}).\boldsymbol{Customer^L_{proc}}$$
$$\boldsymbol{Customer^H} \stackrel{def}{=} (request^H, r_{request}).\boldsymbol{Customer^H_{proc}} + (timeout, r^H_t).\boldsymbol{Customer^H_{preempt}}$$
$$\boldsymbol{Customer^H_{preempt}} \stackrel{def}{=} (preempt, r_{preempt}).\boldsymbol{Customer^H_{proc}} + (timeout, r^H_t).\boldsymbol{Customer^H}$$

After the response, customers wait for the service to finish, notified by the end^L and end^H actions. High priority customers can preempt the service of low priority customers, in which case the low priority service is cancelled and the customer has to restart the whole procedure. Both customers also have to cater for the possibility of the service failing and need to restart

when the server is not responding for a period of time:

$$\mathbf{Customer}^{\mathbf{L}}_{\mathbf{proc}} \stackrel{def}{=} (end^{L}, r_{\mathbf{end}}).\mathbf{Customer}^{\mathbf{L}}_{\mathbf{think}} + (drop, r_{\mathbf{drop}}).\mathbf{Customer}^{\mathbf{L}} + (timeout, r^{B}_{t}).\mathbf{Customer}^{\mathbf{L}}$$
$$\mathbf{Customer}^{\mathbf{H}}_{\mathbf{proc}} \stackrel{def}{=} (end^{H}, r_{\mathbf{end}}).\mathbf{Customer}^{\mathbf{H}}_{\mathbf{think}} + (timeout, r^{B}_{t}).\mathbf{Customer}^{\mathbf{H}}$$

Finally, both customers are allowed to perform their respective think action, which represents the end of that phase of service for that customer:

$$\mathbf{Customer}^{\mathbf{L}}_{\mathbf{think}} \stackrel{def}{=} (think, r_{think}).\mathbf{Customer}^{\mathbf{L}} \qquad \mathbf{Customer}^{\mathbf{H}}_{\mathbf{think}} \stackrel{def}{=} (think, r_{think}).\mathbf{Customer}^{\mathbf{H}}$$

The services can either listen to requests from $\mathbf{H}$ or $\mathbf{L}$ customer and switch between these two states after some time:

$$\mathbf{Service}^{\mathbf{L}} \stackrel{def}{=} (request^{L}, r_{request}).\mathbf{Service}^{\mathbf{L}}_{\mathbf{proc}} + (switch, r^{LH}_{\mathbf{switch}}).\mathbf{Service}^{\mathbf{H}}$$
$$\mathbf{Service}^{\mathbf{H}} \stackrel{def}{=} (request^{H}, r_{request}).\mathbf{Service}^{\mathbf{H}}_{\mathbf{proc}} + (switch, r^{HL}_{\mathbf{switch}}).\mathbf{Service}^{\mathbf{L}}$$

After a request, the services initiate the processing for customers, with a possibility of failing. In addition, services serving low priority customers listen to requests for preemption, which cause the service to drop the served customer and switch to a higher priority customer:

$$\mathbf{Service}^{\mathbf{L}}_{\mathbf{proc}} \stackrel{def}{=} (proc, r^{L}_{\mathbf{proc}}).\mathbf{Service}^{\mathbf{L}}_{\mathbf{end}} + (preempt, r_{preempt}).\mathbf{Service}^{\mathbf{L}}_{\mathbf{drop}} + (break, r_{break}).\mathbf{Service}^{\mathbf{L}}$$
$$\mathbf{Service}^{\mathbf{L}}_{\mathbf{drop}} \stackrel{def}{=} (drop, r_{\mathbf{drop}}).\mathbf{Service}^{\mathbf{H}}_{\mathbf{proc}} + (break, r_{break}).\mathbf{Service}^{\mathbf{L}}$$
$$\mathbf{Service}^{\mathbf{H}}_{\mathbf{proc}} \stackrel{def}{=} (proc, r^{H}_{\mathbf{proc}}).\mathbf{Service}^{\mathbf{H}}_{\mathbf{end}} + (break, r_{break}).\mathbf{Service}^{\mathbf{H}}$$

When finished processing, the servers notify the respective customers:

$$\mathbf{Service}^{\mathbf{L}}_{\mathbf{end}} \stackrel{def}{=} (end^{L}, r_{\mathbf{end}}).\mathbf{Service}^{\mathbf{L}} + (break, r_{break}).\mathbf{Service}^{\mathbf{L}}$$
$$\mathbf{Service}^{\mathbf{H}}_{\mathbf{end}} \stackrel{def}{=} (end^{H}, r_{\mathbf{end}}).\mathbf{Service}^{\mathbf{H}} + (break, r_{break}).\mathbf{Service}^{\mathbf{H}}$$

The whole systems consists of n^{L}_{C} copies of the $\mathbf{Customer}^{\mathbf{L}}$ component and n^{H}_{C} copies of $\mathbf{Customer}^{\mathbf{H}}$, n^{L}_{S} services in the $\mathbf{Service}^{\mathbf{L}}$ and n^{H}_{S} in the $\mathbf{Service}^{\mathbf{H}}$ state:

$$\mathbf{Customers}\{\mathbf{Customer}^{\mathbf{L}}[n^{L}_{C}] \wr\wr \mathbf{Customer}^{\mathbf{H}}[n^{H}_{C}]\} \underset{A}{\bowtie} \mathbf{Services}\{\mathbf{Service}^{\mathbf{L}}[n^{L}_{S}] \wr\wr \mathbf{Service}^{\mathbf{H}}[n^{H}_{S}]\}$$

where the synchronised actions are $A = \{request^{L}, request^{H}, end^{L}, end^{H}, preempt, drop\}$.

It is easy to see that the model is *split-free* and therefore the resulting ODEs approximating moments of component counts are piecewise linear [2]. This allows for the possibility of inexpensive computation of a Lyapunov function as described in Sections 4.2.1 and 4.2.2 which will guarantee convergence of the steady-state passage time as the scale of the system increases provided by Theorem 4.6.

We need to modify the model as the passage times are not explicitly represented, in the same fashion as in the case of the client/server model in Section 3. For example if we are interested in the passage time of a $\mathbf{Customer}^{\mathbf{H}}$ executing one *think* action from the model's initial state, we create a new absorbing copy of the derivative state of $\mathbf{Customer}^{\mathbf{H}}$ and add a transition after the first *think* action.

Fig. 6 shows an excerpt of the GPA source file implementing the following experiments. We use the local cooperation feature of GPA to perform the model modification and conveniently express the absorbing component populations. Each customer is composed with a memory fluid component similar to those described in Section 3.2.1 which remembers whether the first think action has been performed. The initial state `NotFinished`, defined on line 2 in Fig. 6 for $\mathbf{L}$ customers, changes to the absorbing state `Finished` after a cooperation on the think action. The pattern matching feature of GPA then provides a shorthand to express the sum of all absorbing customer components; these are all the components with the memory in the `Finished` state. For convenience the expressions for individual passage time CDFs are stored in variables on lines 7 and 8.

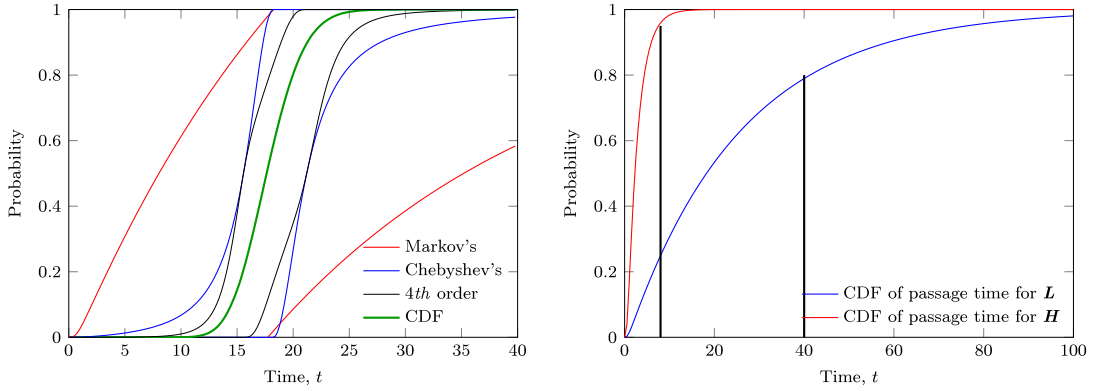
Using the modified model, we first investigate the global passage times from Section 3.1. For example, we can look at the passage time for half of the n^{L}_{C} low priority customers finishing their first *think* action. Fig. 7(a) shows the lower and upper approximations to the CDF of this passage time derived from the Markov and Chebyshev's inequalities, for the system with 100 $\mathbf{L}$ customers, 30 $\mathbf{H}$ customers, 20 services initially for $\mathbf{L}$ and 20 for $\mathbf{H}$.[21] Line 12 in Fig. 6 gives an example of the GPA expression used to produce one of the approximations based on the Chebyshev's inequality. Fig. 7(a) also compares these approximations to the bounds obtained by applying a method of Tari et al. [32] using moments obtained by fluid approximation up to order 4. For approximations based on the Markov inequality, the resulting system of ODEs contains 21 equations. In the case of Chebyshev's inequality, second order moments are needed and the system contains 252 equations. To produce bounds from moments of order up to 4 by the method of Tari, there are 12649 moment ODEs, which took 3 minutes to solve on an Intel Core 2 Duo Linux machine.

The individual passage times from Section 3.2 can be used to express the above SLAs. The steady-state individual passage times from Section 3.2.1 are suitable for the more realistic case where the SLAs are guaranteed only after the system has been

[21] With rate parameters: $r_{request} = 2$, $r^{H}_{t} = 1$, $r_{preempt} = 3$, $r^{L}_{\mathbf{proc}} = 1$, $r^{H}_{\mathbf{proc}} = 2$, $r_{break} = 1$, $r^{B}_{t} = 1$, $r_{\mathbf{end}} = 1$, $r^{LH}_{\mathbf{switch}} = 1$, $r^{HL}_{\mathbf{switch}} = 2$, $r_{think} = 3$.

```
...
NotFinishedL = (think,rthinkL).FinishedL;
FinishedL = (think,rthinkL).FinishedL;

Customers{CustomerL<think>NotFinishedL[n_CL]|CustomerH<think>NotFinishedH[n_CH]}...

$passageCDFL = E[Customers:_<think>FinishedL]/n_CL;
$passageCDFH = E[Customers:_<think>FinishedH]/n_CH;
$sumL = Customers:_<think>FinishedL;

ODEs(stopTime = 100.0, stepSize = 0.1, density = 10){
    Var[$sumL]/(Var[$sumL]+(E[$sumL]-0.5*n_CL)^2);                    //Figure 6a
    $passageCDFL,$passageCDFH;                                        //Figure 6b
}

Iterate rtimeoutH from 0.0 to 5.0 in 50 steps
             n_SL from 10.0 to 60.0 in 50 steps
       ODEs(...) plot {
           $passageCDFH at 8.0 when $passageCDFH at 8.0 >= 0.95;      //Figure 9a
           1.0 when $passageCDFH at  8.0 >= 0.95 and                  //Figure 9b
                    $passageCDFL at 40.0 >= 0.8;
}
```

Fig. 6. Excerpt from the GPA source code used for the experiments.



(a) Half of $\boldsymbol{L}$ finishing. (b) Steady-state individual passage times.

Fig. 7. Figure (a) shows lower and upper approximations to the passage time CDF for half of the low priority customers to finish. The grey lines are the lower and upper bounds obtained from the Tari method [32] using the first 4 moments obtained by fluid approximation. Figure (b) shows the steady-state individual passage times. The thick black lines correspond to the two SLAs; the CDF for $\boldsymbol{L}$ customers just falls short of the 80% requirement that it completes within 40 s.

run for a sufficiently long initial period. Fig. 7(b) shows the steady-state individual passage times for each type of client to finish their respective *think* action in a system with 40 available services.[22] It is produced using the techniques described in Section 3.2.1. In the language of Definition 3.3, these steady-state individual passage times can be specified by the following memory fluid components:

$$\boldsymbol{Mem_S} \stackrel{def}{=} (think, \top).\boldsymbol{Mem_S}$$

and:

$$\boldsymbol{Mem_T} \stackrel{def}{=} (think, \top).\boldsymbol{Mem_{T1}}$$
$$\boldsymbol{Mem_{T1}} \stackrel{def}{=} (think, \top).\boldsymbol{Mem_{T1}}$$

with $\mathcal{S} := \{(\boldsymbol{Mem_S}, \boldsymbol{Mem_S}, think)\}$ and $\mathcal{T} := \boldsymbol{Mem_{T1}}$.

In order to prove convergence of these approximations, we may apply the approaches of Sections 4.2.1 and 4.2.2. The 14 piecewise linear ODEs for the unmodified system are first reduced following the method in Section 4.2.1 which results in a system of 11 ODEs with a single fixed point. Three linear population invariants corresponding to the conservation of

22 Since the two service states can switch between each other, in the steady state the individual passage time does not depend on the proportion of initial counts n_S^L and n_S^H.

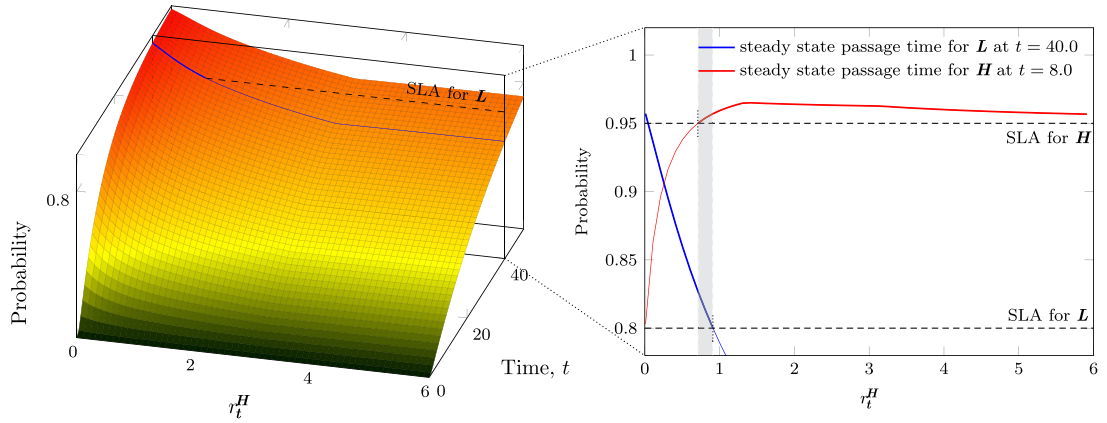
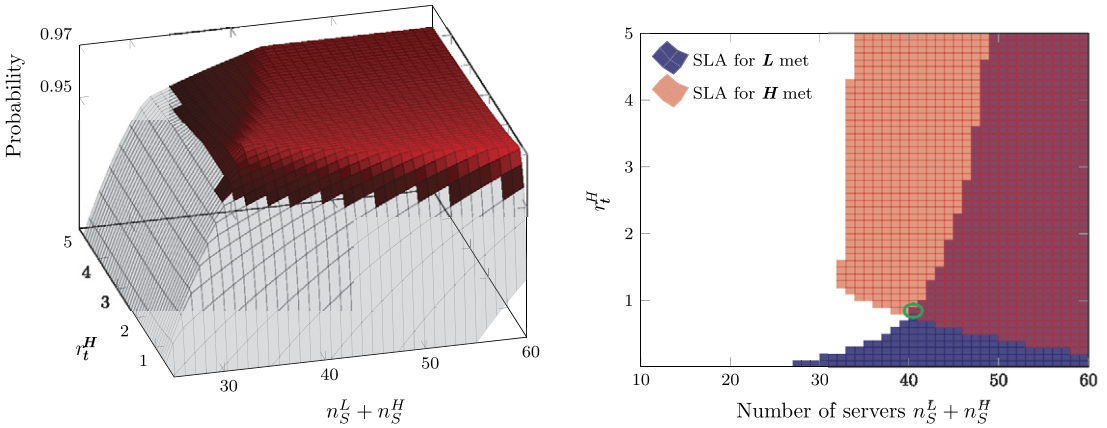


Fig. 8. The effect of the rate r_t^H on the steady state individual passage time probabilities of the customers. The left figure shows the CDF of the individual passage time of $\boldsymbol{L}$ customers for a range of values of the r_t^H rate. The figure on the right compares this with the probability at time $t = 8$ s taken from a similar plot for the $\boldsymbol{H}$ customers.



(a) Individual passage time for the $\boldsymbol{H}$ clients taken at the SLA time of 8 s. The darker shaded surface shows the configurations where the SLA is satisfied.

(b) The intersection of system configurations that satisfy both $\boldsymbol{H}$ client and $\boldsymbol{L}$ client SLAs. The configuration requiring the least number of services is circled.

Fig. 9. The combined effect on the steady state individual passage time SLAs of changing the rate r_t^H and the total number of services, $n_S^L + n_S^H$.

the total population size of the two types of customer and one type of service components are automatically identified by this method. This fixed point is verified to be globally asymptotically stable by finding a feasible solution to a system of 64 LMIs as described in Section 4.2.2. This whole process together with solving the corresponding ODEs is computationally efficient and takes only a few seconds on a standard Intel Core 2 Duo Linux machine. We may then use Theorem 4.6 to obtain confidence in the convergence of these steady-state individual passage-time approximations as the component populations are scaled up.

Fig. 7(b) shows that this particular parameter configuration fails to meet the SLA for $\boldsymbol{L}$ customers. Assuming the communication rates and the server break rates are fixed, the system design offers variability in the scheduling policy: that is, how fast the services switch between serving the two customer classes, $r_{\boldsymbol{switch}}^{HL}$ and $r_{\boldsymbol{switch}}^{LH}$, and how frequently the high priority customers are allowed to request preemptions, r_t^H. A modeller will wish to find a configuration of these parameters that satisfies both SLAs. The left hand side of Fig. 8 shows the steady-state individual passage time CDF for $\boldsymbol{L}$ customers with the fixed SLA deadline of 40 s as the rate r_t^H varies. The range of possible values of r_t^H was discretised into 50 values and the steady-state passage time analysis run for each. The SLA is met only when the probability at time $t = 40$ s, highlighted with the blue curve, is above 0.8, the dashed line. Only small values of r_t^H result in the SLA being satisfied. The right hand side of Fig. 8 compares this with the passage-time CDF of $\boldsymbol{H}$ customers at the SLA deadline of 8 s as the rate r_t^H varies. The SLAs are jointly satisfied only when the rate is within the interval between around 0.7 and 0.9, highlighted in the plot.

Assuming the services have an associated running cost, the provider will be interested in reducing their numbers. Fig. 9(a) shows the passage time probability of $\boldsymbol{H}$ customers taken at the deadline of 8 s as both the rate r_t^H and the numbers of servers $n_S^L + n_S^H$ vary. Only some of the displayed configurations satisfy the SLA requirement that probability of completion be greater than 0.95. Fig. 9(b) shows the configurations for which both the SLA for $\boldsymbol{L}$ customers, at time 40 s, and the SLA for $\boldsymbol{H}$ customers

are satisfied. The non-empty intersection shows all the configurations that the system provider can consider. Out of these, a particular circled configuration minimises the number of running services to 40. If, additionally, the remaining scheduling policy rates r^{LH}_{switch} and r^{HL}_{switch} are allowed to vary, we can sweep to find a parameter combination with only 38 services that still satisfies the two SLAs. The `Iterate` command of GPA was used to explore 2500 different configurations. Line 19 of Fig. 6 shows the command to produce Fig. 9(a) and lines 20–21 show how to produce the intersection of the two feasible regions on Fig. 9(b). Repeating the ODE analysis for all the 2500 configurations takes only around a minute on a standard Intel Core 2 Duo Linux machine.

6. 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 the fluid analysis of large Markov chains to allow us to approximate passage-time distributions. We have introduced the notion of global and individual passage times as being useful quantities in the context of massively-parallel systems and have then gone on to develop systematic methodologies for their approximation with fluid techniques.

We have shown a limiting result for global passage times (Theorem 4.4), 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, for global passage times, we have shown that fluid techniques can establish both upper and lower approximations of the cumulative distribution function of system-wide passage times in the Markov chain (Section 3.1.2).

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 (Theorems 4.5 and 4.6). Furthermore, proving convergence of these passage-time approximations in the steady-state has required us to develop a powerful new approach coupling system reduction techniques with stability verification techniques from control theory [45]. We apply this together with a new continuous-time steady-state convergence result extending similar discrete-time results of Benaïm and Le Boudec [7].

We have 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, for first and second order, be fairly loose, and we have shown that the situation can be improved considerably by considering higher-order moment approximations using the Chebyshev inequality or the Tari method [32] for distribution approximation. 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 lower approximations are directly useful.

In the individual passage time case, the fluid approximation converges relatively quickly to the CDF, making it particularly useful from an engineering perspective. In our case study, we showed that the rapid evaluation of passage time SLAs allowed us to optimise the parameters of the system. Simultaneously we were able to discover configurations of parameters which satisfied potentially several SLAs, thus allowing us to solve a constraint-based scalability problem: in this case, how many servers were required to meet the passage-time service requirements.

Appendix A. Systems of Equations

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

A.1. First-moment differential equations for $\mathbf{CS}(n, m)$

$$
\begin{aligned}
\dot{v}_{\mathbf{C}}(t) &= -\min(v_{\mathbf{C}}(t), v_{\mathbf{S}}(t))r_r + v_{\mathbf{Cw}}(t)r_{tmt} + v_{\mathbf{Ct}}(t)r_t \\
\dot{v}_{\mathbf{Cw}}(t) &= -\min(v_{\mathbf{Cw}}(t), v_{\mathbf{Sg}}(t))r_d - v_{\mathbf{Cw}}(t)r_{tmt} + \min(v_{\mathbf{C}}(t), v_{\mathbf{S}}(t))r_r \\
\dot{v}_{\mathbf{Ct}}(t) &= -v_{\mathbf{Ct}}(t)r_t + \min(v_{\mathbf{Cw}}(t), v_{\mathbf{Sg}}(t))r_d \\
\dot{v}_{\mathbf{S}}(t) &= -\min(v_{\mathbf{C}}(t), v_{\mathbf{S}}(t))r_r - v_{\mathbf{S}}(t)r_b + \min(v_{\mathbf{Cw}}(t), v_{\mathbf{Sg}}(t))r_d + v_{\mathbf{Sb}}(t)r_{rst} \\
\dot{v}_{\mathbf{Sg}}(t) &= -\min(v_{\mathbf{Cw}}(t), v_{\mathbf{Sg}}(t))r_d - v_{\mathbf{Sg}}(t)r_b + \min(v_{\mathbf{C}}(t), v_{\mathbf{S}}(t))r_r \\
\dot{v}_{\mathbf{Sb}}(t) &= -v_{\mathbf{Sb}}(t)r_{rst} + v_{\mathbf{S}}(t)r_b + v_{\mathbf{Sg}}(t)r_b
\end{aligned}
\tag{A.1}
$$

A.2. First-moment differential equations for $\mathbf{CS}'(n, m)$

$$
\begin{aligned}
\dot{v}_{\mathbf{Cw}'}(t) &= -v_{\mathbf{Cw}'}(t)r_g - \frac{v_{\mathbf{Cw}'}(t)}{v_{\mathbf{Cw}'}(t)+v_{\mathbf{Cw}}(t)}\min(v_{\mathbf{Sg}}(t)r_d, v_{\mathbf{Cw}'}(t)r_d + v_{\mathbf{Cw}}(t)r_d) \\
&\quad + \frac{v_{\mathbf{C}'}(t)}{v_{\mathbf{C}'}(t)+v_{\mathbf{C}}(t)}\min(v_{\mathbf{S}}(t)r_r, v_{\mathbf{C}'}(t)r_r + v_{\mathbf{C}}(t)r_r) \\
\dot{v}_{\mathbf{C}'}(t) &= v_{\mathbf{Cw}'}(t)r_g + v_{\mathbf{Ct}}(t)r_s + v_{\mathbf{Ct}'}(t)r_s - \frac{v_{\mathbf{C}'}(t)}{v_{\mathbf{C}'}(t)+v_{\mathbf{C}}(t)}\min(v_{\mathbf{S}}(t)r_r, v_{\mathbf{C}'}(t)r_r + v_{\mathbf{C}}(t)r_r) \\
\dot{v}_{\mathbf{Ct}}(t) &= -v_{\mathbf{Ct}}(t)r_s + \frac{v_{\mathbf{Cw}}(t)}{v_{\mathbf{Cw}'}(t)+v_{\mathbf{Cw}}(t)}\min(v_{\mathbf{Sg}}(t)r_d, v_{\mathbf{Cw}'}(t)r_d + v_{\mathbf{Cw}}(t)r_d) \\
\dot{v}_{\mathbf{Cw}}(t) &= -v_{\mathbf{Cw}}(t)r_g - \frac{v_{\mathbf{Cw}}(t)}{v_{\mathbf{Cw}'}(t)+v_{\mathbf{Cw}}(t)}\min(v_{\mathbf{Sg}}(t)r_d, v_{\mathbf{Cw}'}(t)r_d + v_{\mathbf{Cw}}(t)r_d) \\
&\quad + \frac{v_{\mathbf{C}}(t)}{v_{\mathbf{C}'}(t)+v_{\mathbf{C}}(t)}\min(v_{\mathbf{S}}(t)r_r, v_{\mathbf{C}'}(t)r_r + v_{\mathbf{C}}(t)r_r) \\
\dot{v}_{\mathbf{C}}(t) &= v_{\mathbf{Cw}}(t)r_g - \frac{v_{\mathbf{C}}(t)}{v_{\mathbf{C}'}(t)+v_{\mathbf{C}}(t)}\min(v_{\mathbf{S}}(t)r_r, v_{\mathbf{C}'}(t)r_r + v_{\mathbf{C}}(t)r_r) \\
\dot{v}_{\mathbf{Ct}'}(t) &= -v_{\mathbf{Ct}'}(t)r_s + \frac{v_{\mathbf{Cw}'}(t)}{v_{\mathbf{Cw}'}(t)+v_{\mathbf{Cw}}(t)}\min(v_{\mathbf{Sg}}(t)r_d, v_{\mathbf{Cw}'}(t)r_d + v_{\mathbf{Cw}}(t)r_d) \\
\dot{v}_{\mathbf{Sb}}(t) &= v_{\mathbf{Sg}}(t)r_b - v_{\mathbf{Sb}}(t)r_f + v_{\mathbf{S}}(t)r_b \\
\dot{v}_{\mathbf{Sg}}(t) &= -v_{\mathbf{Sg}}(t)r_b - \min(v_{\mathbf{Sg}}(t)r_d, v_{\mathbf{Cw}'}(t)r_d + v_{\mathbf{Cw}}(t)r_d) + \min(v_{\mathbf{S}}(t)r_r, v_{\mathbf{C}'}(t)r_r + v_{\mathbf{C}}(t)r_r) \\
\dot{v}_{\mathbf{S}}(t) &= v_{\mathbf{Sb}}(t)r_f - v_{\mathbf{S}}(t)r_b + \min(v_{\mathbf{Sg}}(t)r_d, v_{\mathbf{Cw}'}(t)r_d + v_{\mathbf{Cw}}(t)r_d) - \min(v_{\mathbf{S}}(t)r_r, v_{\mathbf{C}'}(t)r_r + v_{\mathbf{C}}(t)r_r)
\end{aligned}
$$

Appendix B. Proofs and lemmas

B.1. Lipschitz continuity of ODEs

For any GPEPA model, the following lemma verifies Lipschitz continuity of $\mathbf{f} : \mathbb{R}_+^{\mathcal{N}(G)} \to \mathbb{R}^{\mathcal{N}(G)}$ defined in Section 2. It uses the technical notion of the *structural depth* of a GPEPA model. This is the largest number of cooperations involving action type α, whose immediate effect can be seen by a fluid component enabling an α-action within some component group. The formal definition follows the lemma below. Furthermore, we will require the *maximal local alpha-rate* for an action type α defined by $\mathcal{Q}_\alpha^{\max}(G) := \max_{(H,P)\in\mathcal{B}(G)}\{r_\alpha(P)\}$ and the *maximal local rate* defined by $\mathcal{Q}^{\max}(G) := \max_{(H,P)\in\mathcal{B}(G)}\{\sum_{\alpha\in\mathcal{A}} r_\alpha(P)\}$.

Lemma B.1. *The system of ODEs, $\dot{\mathbf{v}}(t) = \mathbf{f}(\mathbf{v}(t))$, corresponding to a GPEPA 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. We see from Definition 2.4 that for $1 \le k \le \mathcal{N}(G)$ (corresponding to some $(H, P) \in \mathcal{B}(G)$) and $\mathbf{v} \in \mathbb{R}_+^{\mathcal{N}(G)}$:

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

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 a straightforward application of structural induction over Definition 2.1 to see that it has the following general form, for some $1 \le i \le \mathcal{N}(G)$ and $1 \le D \le \mathcal{D}_\alpha(G)$:

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

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

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

Now, the functions $a_n(\cdot)$ and $b_n(\cdot)$ are just instances of apparent rate (Definition 2.2). 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 Eq. (B.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:

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

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

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

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:

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

Applying the inequalities of Eq. (B.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 (Definition 2.2) 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 fluid 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:

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

We now apply Lemma B.3 on the open convex set, int(A) to show that $\mathbf{f}$ is Lipschitz continuous on A with a Lipschitz constant $\mathcal{K}(G) := 2\mathcal{N}(G) \sum_{\alpha \in \mathcal{A}} (\mathcal{D}_\alpha(G) + 1)\mathcal{Q}_\alpha^{\max}(G)$.

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. □

Definition B.2 (*Structural Depth*)**.** For any GPEPA model G and action type $\alpha \in \mathcal{A}$, the structural depth of G with respect to α is $\mathcal{D}_\alpha(G)$, defined as follows.

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

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

Lemma B.3. *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:

$$\begin{aligned} |G_i(1) - G_i(0)|^2 &= \left[\int_0^1 \frac{\mathrm{d}G_i}{\mathrm{d}t}(s)\, \mathrm{d}s \right]^2 = \left[\sum_{j=1}^n (y_j - x_j) \right]^2 \left[\int_0^1 \frac{\partial g_i}{\partial x_j}((1-s)\mathbf{x} + s\mathbf{y})\, \mathrm{d}s \right]^2 \\ &\leq n \|\mathbf{y} - \mathbf{x}\|^2 \Lambda^2 \end{aligned}$$

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. □

References

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

[2] R. Hayden, J.T. Bradley, A fluid analysis framework for a Markovian process algebra, Theoretical Computer Science 411 (May 2010) 2260–2297.

[3] J. Hillston, Fluid flow approximation of PEPA models, in: Proceedings of the Second International Conference on the Quantitative Evaluation of Systems, Torino, Italy, IEEE Computer Society Press, September 2005, pp. 33–43.

[4] L. Bortolussi, A. Policriti, Stochastic concurrent constraint programming and differential equations, in: QAPL'07, 5th Workshop on Quantitative Aspects of Programming Languages, in: Electronic Notes in Theoretical Computer Science, vol. 190, September 2007, pp. 27–42.

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

[6] R. David, H. Alla, Discrete, Continuous, and Hybrid Petri Nets, 2nd ed., Springer, Berlin Heidelberg, 2010.

[7] M. Benaïm, J.-Y. Le Boudec, A class of mean field interaction models for computer and communication systems, Performance Evaluation 65 (11–12) (2008) 823–838.

[8] A. Bobbio, M. Gribaudo, M. Telek, Analysis of Large Scale Interacting Systems by Mean Field Method, in: Fifth International Conference on Quantitative Evaluation of Systems, IEEE, September 2008, pp. 215–224.

[9] R. Bakhshi, L. Cloth, W. Fokkink, B.R. Haverkort, Mean-field framework for performance evaluation of push-pull gossip protocols, Performance Evaluation 68 (February 2011) 157–179.

[10] M. Tribastone, Scalable analysis of stochastic process algebra models, Ph.D. Thesis, University of Endinburgh, 2010.

[11] J. Little, A proof for the queuing formula: $L = \lambda W$, Operations Research 9 (3) (1961) 383–387.

[12] J. Ding, Structural and fluid analysis for large scale PEPA models — with applications to content adaptation systems, Ph.D. Thesis, The University of Edinburgh, 2010.

[13] A. Clark, A. Duguid, S.T. Gilmore, M. Tribastone, Partial evaluation of PEPA models for fluid-flow analysis, in: N. Thomas, C. Juiz (Eds.), Proceedings of the 5th European Performance Engineering Workshop on Computer Performance Engineering, in: Lecture Notes in Computer Science (Berlin, Heidelberg), vol. 5261, Springer, Berlin Heidelberg, August 2008, pp. 2–16.

[14] G. Kesidis, T. Konstantopoulos, P. Sousi, A stochastic epidemiological model and a deterministic limit for BitTorrent-like peer-to-peer file-sharing networks, in: 2nd Workshop on Network Control and Optimization, NET-COOP, November 2008, pp. 26–36.

[15] J.T. Bradley, R. Hayden, W.J. Knottenbelt, T. Suto, Extracting response times from fluid analysis of performance models, in: SIPEW'08, SPEC International Performance Evaluation Workshop, Darmstadt, 27–28 June 2008, in: Lecture Notes in Computer Science, vol. 5119, May 2008, pp. 29–43.

[16] N.J. Dingle, W.J. Knottenbelt, Automated customer-centric performance analysis of generalised stochastic petri nets using tagged tokens, in: Third International Workshop on the Practical Application of Stochastic Modelling (PASM), vol. 232, Amsterdam, Elsevier Science Publishers B.V., March 2009, pp. 75–88.

[17] G. Balbo, M. Beccuti, M. De Pierro, G. Franceschinis, First passage time computation in tagged GSPNs with queue places, The Computer Journal, July 2010, doi:10.1093/comjnl/bxq056.

[18] A. Stefanek, R.A. Hayden, J.T. Bradley, A new tool for the performance analysis of massively parallel computer systems, in: QAPL'10, 8th Workshop on Quantitative Aspects of Programming Languages, Electronic Proceedings of Theoretical Computer Science, vol. 28, 2010, pp. 159–181.

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

[20] S. Gilmore, J. Hillston, M. Ribaudo, An efficient algorithm for aggregating PEPA models, IEEE Transactions on Software Engineering 27 (May 2001) 449–464.

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

[22] N. Geisweiller, J. Hillston, M. Stenico, Relating continuous and discrete PEPA models of signalling pathways, Theoretical Computer Science 404 (November 2008) 97–111.
[23] J. Júlvez, E. Jiménez, L. Recalde, M. Silva, On observability in timed continuous Petri net systems, in: QEST'04, 1st International Conference on Quantitative Evaluation of Systems, vol. 266, IEEE, September 2004, pp. 60–69.
[24] R.K. Miller, A.N. Michel, Ordinary Differential Equations, Academic Press, 1982.
[25] C.S. Gillespie, Moment-closure approximations for mass-action models, IET Systems Biology 3 (January 2009) 52–58.
[26] C.H. Lee, K.-H. Kim, P. Kim, A moment closure method for stochastic reaction networks, The Journal of Chemical Physics 130 (April 2009) pp. 134107–134107–15.
[27] J. Hespanha, Moment closure for biochemical networks, in: Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on, March 2008, pp. 142 –147.
[28] N.J. Dingle, P.G. Harrison, 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 64 (August 2004) 908–920.
[29] C. Baier, B.R. Haverkort, H. Hermanns, J.-P. Katoen, Model checking continuous-time Markov chains by transient analysis, in: 12th International Conference on Computer Aided Verification, July 2000, pp. 358–372.
[30] A. Shwartz, A. Weiss, Large Deviations for Performance Analysis, Chapman and Hall, 1995.
[31] R.A. Hayden, Ph.D. Thesis additional material. http://rhayden.me.uk/thesis, 2011.
[32] A. Tari, M. Telek, P. Buchholz, A unified approach to the moments based distribution estimation — unbounded support, in: M. Bravetti, L. Kloul, G. Zavattaro (Eds.), EPEW 2005, Formal Techniques for Computer Systems and Business Processes, in: Lecture Notes in Computer Science, vol. 3670, Springer, 2005, pp. 79–93.
[33] J.-Y. Le Boudec, Performance Evaluation of Computer and Communication Systems, EPFL Press, Lausanne, Switzerland, 2010.
[34] R. Serfozo, Basics of Applied Stochastic Processes, Springer, Berlin Heidelberg, 2009.
[35] M. Tribastone, S.T. Gilmore, J. Hillston, Scalable differential analysis of process algebra models, IEEE Transactions on Software Engineering PP (99) (2010) 1.
[36] T. Kurtz, Solutions of Ordinary Differential Equations as Limits of Pure Jump Markov Processes, vol. 7, April 1970, pp. 49–58.
[37] R.W.R. Darling, J.R. Norris, Differential equation approximations for Markov chains, Probability Surveys 5 (2008) 37.
[38] M. Benaïm, Recursive algorithms, urn processes and chaining number of chain recurrent sets, Ergodic Theory and Dynamical Systems 18 (1) (1998) 53–87.
[39] T. Kurtz, S. Ethier, Markov Processes Characterisation and Convergence, Wiley, 1986.
[40] R.A. Hayden, Scalable performance analysis of massively parallel stochastic systems. Ph.D. Thesis, Imperial College London, 2011. http://pubs.doc.ic.ac.uk/hayden-thesis.
[41] J.K. Scott, P.I. Barton, Tight, efficient bounds on the solutions of chemical kinetics models, Computers & Chemical Engineering 34 (May 2010) 717–731.
[42] M. Silva, E. Terue, J. Colom, Linear algebraic and linear programming techniques for the analysis of place/transition net systems, in: Lecture Notes in Computer Science, vol. 1491, Springer, Berlin/Heidelberg, 1998, pp. 309–373.
[43] A. Ben-Israel, T.N.E. Greville, Generalized Inverses: Theory and Applications, 2nd ed., Springer, 2003.
[44] H.K. Khalil, Nonlinear Systems, 3rd ed., Prentice Hall, 2002.
[45] A. Pavlov, N. van de Wouw, H. Nijmeijer, Convergent piecewise affine systems: analysis and design, Part I: continuous case, in: 44th IEEE Conference on Decision and Control, CDC-ECC '05, December 2005, pp. 5391–5396.
[46] S. Boyd, L. El Ghaoui, E. Feron, V. Balakrishnan, Linear matrix inequalities in system and control theory. Society for Industrial and Applied Mathematics, 1994.
[47] The MathWorks, Introduction to linear matrix inequalities (Robust Control Toolbox™). http://www.mathworks.com/access/helpdesk/help/toolbox/robust/ug/f8-2138.html.
[48] M. Johansson, A. Rantzer, Computation of piecewise quadratic Lyapunov functions for hybrid systems, IEEE Transactions on Automatic Control 43 (April 1998) 555–559.
[49] M. Brin, G. Stuck, Introduction to Dynamical Systems, Cambridge University Press, 2002.
[50] P. Billingsley, Convergence of Probability Measures, John Wiley & Sons, 1968.
[51] O. Kallenberg, Foundations of Modern Probability, Springer, 2002.
[52] L.C.G. Rogers, D. Williams, Diffusions, Markov Processes and Martingales, Volume 1: Foundations, Cambridge University Press, 2000.