Noname manuscript No.
(will be inserted by the editor)

A computational approach to steady-state convergence of fluid limits for Coxian queuing networks with abandonment

Max Tschaikowski · Mirco Tribastone

Received: date / Accepted: date

Abstract Many-server queuing networks with general service and abandonment times have proven to be a realistic model for scenarios such as call centers and health-care systems. The presence of abandonment makes analytical treatment difficult for general topologies. Hence, such networks are usually studied by means of fluid limits. The current state of the art, however, suffers from two drawbacks. First, convergence to a fluid limit has been established only for the transient, but not for the steady state regime. Second, in the case of general distributed service and abandonment times, convergence to a fluid limit has been either established only for a single queue, or has been given by means of a system of coupled integral equations which does not allow for a numerical solution. By making the mild assumption of Coxian-distributed service and abandonment times, in this paper we address both drawbacks by establishing convergence in probability to a system of coupled ordinary differential equations (ODEs) using the theory of Kurtz. The presence of abandonments leads in many cases to ODE systems with a global attractor, which is known to be a sufficient condition for the fluid and the stochastic steady state to coincide in the limiting regime. The fact that our ODE systems are piecewise affine enables a computational method for establishing the presence of a global attractor, based on a solution of a system of linear matrix inequalities.

Keywords Abandonment Queueing Networks · Fluid Limits · Linear Matrix Inequalities

1 Introduction

Like all modeling techniques with a discrete state space representation, queuing networks with an underlying dynamics based on continuous-time Markov chains (CTMCs) suffer from the well-known problem of state space explosion. This makes large-scale models with many jobs and servers very difficult to analyze in practice, since it leads to unfeasibly large matrices in the case of numerical solutions [35] or to long runtimes when simulation is

Max Tschaikowski
School for Advanced Studies Lucca, Italy
E-mail: max.tschaikowski@imtlucca.it

Mirco Tribastone
School for Advanced Studies Lucca, Italy
E-mail: mirco.tribastone@imtlucca.it

used. There is a vast literature concerned with techniques developed to tackle this issue, based for instance on the exploitation of the presence of a product form for the stationary distribution [3].

Another approach, which is the focus of this paper, is represented by studying *fluid limits* for Markov population processes by means of ordinary differential equations (ODEs) in the sense of Kurtz [20], known in the area of queueing networks also as the Halfin-Whitt regime [14]. In this context, it is possible to identify a sequence of CTMCs characterized by the same network topology but with increasingly larger job and server sizes that converge in probability to the solution of the ODE system. In practice, this can be seen as an approximation technique where the CTMC's state space is approximated by a continuous state-space dynamics that estimates the average queue lengths in the network.

While the underlying ODE system can be solved much more efficient than the stochastic process using numerical integration [2], a drawback of this approach is that convergence holds for finite time intervals only. However, due to the theoretical and practical importance of analyzing models under stationary conditions, there has been considerable interest in studying when convergence can also be extended to the steady state. A crucial result, rooted in Poincare's recurrence theorem, states that the presence of a global attractor for the limit ODE system ensures convergence [13,38]. This fact has been used to derive steady state results for different models in the literature, including models of virtualized environments [1], grid computing [13], garbage collection algorithms [37], and optical packet switches [38].

In this paper we propose a computational method to study steady-state convergence in the fluid limit for a class of queuing networks. In particular, we study many-server networks with generally distributed service and abandonment times. We motivate this choice by the fact that this class has proven to be a realistic model for real-world scenarios such as call centers [12] and health-care systems [30], where abandonment is related to customer's impatience, and in certain computing systems where instead it can describe the occurrence of timeouts due, e.g., to network delays [36]. To keep the overall model a Markov population process, we assume that service- and abandonment-time distributions are Coxian, which allows us to model a general distribution with arbitrary accuracy (e.g., [8,6]). Moreover important, the limit ODE system has a piecewise affine vector field. This enables the use of the theory of linear matrix inequalities (LMI) to conclude that the feasibility of an LMI problem, which admits a computational treatment, guarantees the presence of a global attractor [33]. In case of exponentially distributed service and abandonment times, we additionally allow the abandonment distribution to depend on the fact whether a customer is waiting or is being served. This generalizes the common abandonment policy [28,29] where customers cannot abandon while being served. As discussed, the feasibility of an LMI problem implies our desired result of convergence. A positive side effect of a global attractor is that a solution of the transient regime provides one also with the global attractor itself if the time interval is chosen long enough.

Our approach is generic and thus can be in principle applied to networks with arbitrary topologies and Coxian-distributed service and abandonment times. However, the number of queuing stations and the number of Coxian stages lead to a growth in the number of affine modes of the ODE vector field. This, in turn, impacts on the computational complexity of the LMI feasibility problem. Therefore, we complete our study with an empirical evaluation of our approach using a Matlab implementation. On a set of randomly generated networks, we show that the LMI problem is feasible in a large fraction of cases for networks with exponentially distributed abandonments, albeit with a tendency to degrade as the number of stations or the number of stages grow. In particular, we establish that the LMI approach

applies also to the common policy where customers cannot abandon while being served [28, 29].

Related work. Due to the lack of closed form solutions for queueing systems with general service, one usually resorts to efficient numerical algorithms [32] or to fluid and diffusion models [7]. The situation is similar in the case of abandonment networks, where no analytical solutions are known even if the service, arrival and abandonment times of those are assumed to be exponential [28].

Fluid models are rooted in the functional law of large number and are deterministic, while diffusion models are based on the functional central limit theorem and depend on higher order moments of the service, arrival and abandonment times [10]. Although there exists a considerable amount of literature on diffusion models of many-server queuing networks and limiting regimes of single-server queues (see [10,17,9] and the references therein for more on this topic), in the remainder we relate our work to fluid models of many-server queuing networks, as they are the closest to the topics of this paper.

Mathematically, fluid models are in general described by means of coupled integral equations and, in the classic setting, they approximate the number of customers present in the queues of a network; see, for instance, [28,29] which covers Markovian many-server time-dependent queuing networks with abandonment. In contrast to [28,29], in our model a customer can abandon while being served. In the case of call centers, this can be motivated by the fact that a customer has an appointment at a certain time, while in performance related models customers are jobs/files and abandonment times are timeouts. Because of this, our fluid limits are not comparable to those in [28] in general. In order to cope with general service-, arrival- and abandonment-time distributions, one has to keep track of the waiting times of each customer in queue as well as of the amount of time each customer has been in service. This idea has been used in [39,24,26] to establish a fluid limit result for a single queue with general service and abandonment times and yields a deterministic model than can be solved efficiently. Among stimulating further results for the single queue case [27, 23,22], this works were pivotal for [25,21] in which a fluid model for a time-dependent general network with abandonment and the underlying numerical algorithm for its solution were considered. In contrast to [24], however, [25,21] do not establish that the fluid model is a limit of the stochastic queuing network. By considering a slightly different state descriptor, [19,18] tackle this problem by using a measure valued stochastic process. They are able to establish a fluid limit which is characterized by a set of coupled integral equations. In contrast to [25,21], the network is time-homogenous but the routing of internal and external customers may be different. Unfortunately, the characterization of the deterministic model [18] is implicit and does not allow for a numerical solution; also, it is not clear whether [18] can be used to prove the convergence in [25], due to the differences in the state descriptors.

Paper organization. This paper is structured as follows. Section 2 introduces our reference CTMC model of an open queuing with many-server queues with Coxian distributed service and abandonment times. Section 3 develops the fluid limits. We consider the following cases separately. First we present the case of exponential distributions for both service and abandonment, in Section 3.1; this represents the simplest case, and allows us to clearly point out through an example that the feature of abandonment is necessary for our proposed computational approach to yield feasible LMI problems. Then, Section 3.2 considers the case of Coxian-distributed abandonments, while Section 3.3 discusses the case of Coxian-distributed service and abandonment times. This order allows for a concise presentation.

For each case we discuss their computational complexity which we measure in terms of number of affine modes of the vector field and matrix sizes. Section 4 presents the main results concerning the convergence to the fluid limit, while Section 5 shows the results of the numerical evaluation. In Section 6 we extend our exponential model to the case where customer abandonment times in service may differ from the abandonment times of customers that are waiting. By carrying out a numerical investigation similar to that of Section 5, we are able to relate our abandonment policy to the common one [28,29]. Finally, conclusions are drawn in Section 7.

2 Stochastic Model

Here we describe the underlying CTMC of an open queuing network with many-server stations with Coxian-distributed service and abandonment times. To each station we associate exactly one queue with infinite capacity. We assume the FCFS policy and that a customer can abandon while being served. In the case of call centers, this can be motivated by the fact that a customer has an appointment that he or she cannot miss, while in performance related models timeouts may not be related to the current work progress. Let $n \geq 1$ denote the number of stations and $(r_{i,j})_{1\leq i,j\leq n}$ such that $r_{i,j}$ denotes the probability that a customer joins queue j after being served at station i. We assume that any two stations of the network are connected by a path whose probability is nonzero. Let $\Gamma^i \geq 0$ denote the intensity of the independent Poisson arrival process associated to station i (where $\Gamma^i = 0$ means that station i has no arrivals) and let $k_i \geq 1$ be the number of stages of the Coxian-distributed service times. The corresponding service rates are given by a vector $\boldsymbol{\mu}^i = (\mu^i_k)_{1\leq k\leq k_i}$, where $\mu^i_k \in (0;\infty)$ is the service rate at stage k, and a vector $\mathbf{p}^i = (p^i_k)_{1\leq k\leq k_i-1}$ where $p^i_k > 0$ is the probability with which a job goes into stage $k+1$ after receiving service at stage k. Naturally, $1-p^i_k$ is the probability with which service is completed when at stage k; a job leaves the station from stage k_i with probability 1. In order to simplify later expressions, we define $p^i_{k_i} := 0$. Similarly, the Coxian-distributed abandonment of station i has $l_i \geq 1$ stages and the underlying parameters are $\boldsymbol{\lambda}^i = (\lambda^i_l)_{1\leq l\leq l_i}$, $\mathbf{q}^i = (q^i_l)_{1\leq l\leq l_i-1}$ and $q^i_{l_i} := 0$.

The state of station i is fully characterized by the vector $(X_{S^i_1}, X_{C^i_{k,l}})_{1\leq k\leq k_i, 1\leq l\leq l_i}$ where each element is a nonnegative integer denoting the following:

- $X_{C^i_{1,l}}$ is the population of jobs which are in abandonment-stage l and that are either waiting for service or which are in service-stage 1;
- $X_{S^i_1}$ is the population of servers which are either idle or servicing jobs in service-stage 1 of *any* abandonment-stage l. That is, the servers $X_{S^i_1}$ are shared by $X_{C^i_{1,1}}, \ldots, X_{C^i_{1,l_i}}$;
- $X_{C^i_{k,l}}$, with $k > 1$ and $1 \leq l \leq l_i$, is the population of jobs in service-stage k and abandonment-stage l which are served.

From the above state description one readily infers that the number of jobs and servers in station i is given by $\sum_{k=1}^{k_i}\sum_{l=1}^{l_i} X_{C^i_{k,l}}$ and $X_{S^i_1} + \sum_{k=2}^{k_i}\sum_{l=1}^{l_i} X_{C^i_{k,l}}$, respectively. In particular, the network state descriptor is a vector from $\mathbb{N}_0^{\sum_i (k_i l_i + 1)}$ in the form

$$\mathbf{X} = (X_{S^i_1}, X_{C^i_{k,l}})_{1\leq i\leq n, 1\leq k\leq k_i, 1\leq l\leq l_i}.$$

We now define the transition rates of the CTMC, in the customary form of jump vectors and associated transition functions from a generic state $\mathbf{X}$. We denote by $q(\mathbf{X}, \mathbf{X'})$ the

transition rate from state $\mathbf{X}$ to state $\mathbf{X'}$. Consider some station $1 \leq i \leq n$ and define, for all $1 \leq l \leq l_i$ and $1 \leq j \leq n$,

$$\mathbf{X} + \mathbf{h}^{i,i}_{1,l} = (X_{C^i_{1,l}} - 1, X_{S^i_1} - 1, X_{C^i_{2,l}} + 1, \ldots)\,, \tag{1}$$

$$\mathbf{X} + \mathbf{h}^{i,j}_{1,l} = (X_{C^i_{1,l}} - 1, X_{C^j_{1,1}} + 1, \ldots)\,, \qquad \text{for all} \quad j \neq i, \tag{2}$$

$$\mathbf{X} + \mathbf{h}^{i,i}_{k,l} = (X_{C^i_{k,l}} - 1, X_{C^i_{k+1,l}} + 1, \ldots)\,, \qquad \text{for all} \quad 2 \leq k \leq k_i - 1, \tag{3}$$

$$\mathbf{X} + \mathbf{h}^{i,j}_{k,l} = (X_{C^i_{k,l}} - 1, X_{S^i_1} + 1, X_{C^j_{1,1}} + 1, \ldots)\,, \qquad \text{for all} \quad 2 \leq k \leq k_i, \tag{4}$$

where we use ellipsis to denote all elements of $\mathbf{X}$ which are not affected by a jump. Jump (1) describes a job in service-stage 1 which moves into service-stage 2; (2) describes a job in service-stage 1 which completes service and moves into another station j; (3) denotes a job in service-stage k which moves to the next stage, whereas with (4) the jobs complete service in service-stage k of the Coxian, with the job moving to station j and a server unit becoming available again at service-stage 1 for a new service.

This defines the jumps induced by the service. For the abandonment and arrivals, instead, we define for all $1 \leq i \leq n$

$$\mathbf{X} + \mathbf{h}^{i,0}_{k,l} = (X_{C^i_{k,l}} - 1, X_{C^i_{k,l+1}} + 1, \ldots)\,, \quad \text{for all} \quad 1 \leq k \leq k_i,\ 1 \leq l \leq l_i - 1, \tag{5}$$

$$\mathbf{X} + \mathbf{h}^{i,-}_{1,l} = (X_{C^i_{1,l}} - 1, \ldots)\,, \qquad \text{for all} \quad 1 \leq l \leq l_i, \tag{6}$$

$$\mathbf{X} + \mathbf{h}^{i,-}_{k,l} = (X_{C^i_{k,l}} - 1, X_{S^i_1} + 1, \ldots)\,, \qquad \text{for all} \quad 2 \leq k \leq k_i,\ 1 \leq l \leq l_i, \tag{7}$$

$$\mathbf{X} + \mathbf{h}^{i,+} = (X_{C^i_{1,1}} + 1, \ldots), \tag{8}$$

where jump (5) describes a job which is moving from abandonment-stage l to abandonment-stage $l+1$ and (6) - (7) express the fact that a job leaves the queuing network in abandonment-stage l; finally, (8) describes the arrival of a new job to the network.

According to this description and notation, the transition rates from any state $\mathbf{X}$ are, for all $1 \leq i, j \leq n$, as follows:

$$q(\mathbf{X}, \mathbf{X}+\mathbf{h}^{i,i}_{1,l}) = f(\mathbf{X}, \mathbf{h}^{i,i}_{1,l}) := p^i_1 \mu^i_1 \max\big(\min\big(X_{C^i_{1,l}}, X_{S^i_1} - \sum_{l+1 \leq l' \leq l_i} X_{C^i_{1,l'}}\big), 0\big), \tag{9}$$

$$q(\mathbf{X}, \mathbf{X}+\mathbf{h}^{i,j}_{1,l}) = f(\mathbf{X}, \mathbf{h}^{i,j}_{1,l}) := r_{i,j}(1 - p^i_1)\mu^i_1 \max\big(\min\big(X_{C^i_{1,l}}, X_{S^i_1} - \sum_{l+1 \leq l' \leq l_i} X_{C^i_{1,l'}}\big), 0\big)$$

for $j \neq i$; moreover, we have

$$\begin{aligned}
q(\mathbf{X}, \mathbf{X} + \mathbf{h}^{i,i}_{k,l}) &= f(\mathbf{X}, \mathbf{h}^{i,i}_{k,l}) := p^i_k \mu^i_k X_{C^i_{k,l}}, \\
q(\mathbf{X}, \mathbf{X} + \mathbf{h}^{i,j}_{k,l}) &= f(\mathbf{X}, \mathbf{h}^{i,j}_{k,l}) := r_{i,j}(1 - p^i_k)\mu^i_k X_{C^i_{k,l}}, \\
q(\mathbf{X}, \mathbf{X} + \mathbf{h}^{i,0}_{k,l}) &= f(\mathbf{X}, \mathbf{h}^{i,0}_{k,l}) := q^i_l \lambda^i_l X_{C^i_{k,l}}, \\
q(\mathbf{X}, \mathbf{X} + \mathbf{h}^{i,-}_{k,l}) &= f(\mathbf{X}, \mathbf{h}^{i,-}_{k,l}) := (1 - q^i_l)\lambda^i_l X_{C^i_{k,l}}, \\
q(\mathbf{X}, \mathbf{X} + \mathbf{h}^{i,+}) &= f(\mathbf{X}, \mathbf{h}^{i,+}) := \Gamma^i,
\end{aligned}$$

Apart from the service rates of the service-stage 1 given in (9), all definitions are straightforward. To shade more light on the former, let us define

$$\tilde{l} := \min\Big\{1 \leq l \leq l_i + 1 \mid \sum_{l'=l}^{l_i} X_{C^i_{1,l'}} \leq X_{S^i_1}\Big\}$$

and observe that the service rate of service-stage 1 of abandonment-stage l satisfies

$$\mu_1^i \max\Big(\min\Big(X_{C_{1,l}^i}, X_{S_1^i} - \sum_{l'=l+1}^{l_i} X_{C_{1,l'}^i}\Big), 0\Big) = \\ = \begin{cases} 0 & , \ 1 \leq l < \tilde{l} - 1 \\ \mu_1^i \Big(X_{S_1^i} - \sum_{l'=\tilde{l}}^{l_i} X_{C_{1,l'}^i}\Big) & , \ l = \tilde{l} - 1 \\ \mu_1^i X_{C_{1,l}^i} & , \ \tilde{l} \leq l \leq l_i \end{cases}$$

From this we infer that jobs in abandonment-stage l are served after those in abandonment-stage l' with $1 \leq l < l' \leq l_i$. For instance, let us assume that we are given $k_i = 1$ and $l_i = 3$ with $X_{C_{1,1}^i} = 2$, $X_{C_{1,2}^i} = 3$, $X_{C_{1,3}^i} = 4$ and $X_{S_1^i} = 5$. Then, the service rate of $X_{C_{1,3}^i}$ is $\mu_1^i \max(\min(X_{C_{1,3}^i}, S_1^i)) = 4\mu_1^i$, the service rate of $X_{C_{1,2}^i}$ is $\mu_1^i \max(\min(X_{C_{1,2}^i}, S_1^i - X_{C_{1,3}^i})) = \mu_1^i$ and the service rate of $X_{C_{1,1}^i}$ is $\mu_1^i \max(\min(X_{C_{1,1}^i}, X_{S_1^i} - X_{C_{1,2}^i} - X_{C_{1,3}^i})) = 0$. That is, the 5 servers available at station i serve all 4 jobs of abandonment-stage 3, one of the 3 jobs of abandonment-stage 2 and none of the 2 jobs of abandonment-stage 1.

The CTMC is completely characterized by above transitions and an initial condition

$$\mathbf{X}(0) = \big(X_{S_1^i}(0), X_{C_{k,l}^i}(0)\big)_{1 \leq i \leq n, 1 \leq k \leq k_i, 1 \leq l \leq l_i}.$$

We denote the CTMC by $(\mathbf{X}(t))_{t \geq 0}$, where

$$\mathbf{X}(t) = \big(X_{S_1^i}(t), X_{C_{k,l}^i}(t)\big)_{1 \leq i \leq n, 1 \leq k \leq k_i, 1 \leq l \leq l_i}.$$

Although general initial conditions can be chosen, we require in the sequel that no jobs are served in Coxian service-stages greater than 1 at time point zero. This is without loss of generality but simplifies the exposition, because it ensures that the number of servers in station i satisfies $X_{S_1^i}(t) + \sum_{k=2}^{k_i} \sum_{l=1}^{l_i} X_{C_{k,l}^i}(t) = X_{S_1^i}(0)$ for all $t \geq 0$, see also Remark 1 and the following discussion.

Let us remark that our model can also cover infinite-server (i.e., delay) stations, by setting $S_1^i(0) := \infty$. (Formally, one has to remove S_1^i from the state descriptor and resolve the minima in the equation (9).)

3 Fluid Limit

We are now ready to develop the fluid limit according to Kurtz [20]. Given a model, we consider a vector of initial *densities* in the form

$$\boldsymbol{\chi}(0) = \big(S_1^i(0), C_{k,l}^i(0)\big)_{1 \leq i \leq n, 1 \leq k \leq k_i, 1 \leq l \leq l_i}$$

with $C_{k,l}^i(0) = 0$ for all $2 \leq k \leq k_i$, $C_{1,l}^i(0) \geq 0$ and $S_1^i(0) > 0$. Using this initial vector, we construct a family of CTMCs, $\{(\mathbf{X}_N(t))_{t \geq 0} \mid N \geq 1\}$, by setting the initial state of the N-th CTMC to be $\mathbf{X}_N(0) = \lfloor N \cdot \chi(0) \rfloor$ with probability 1, where $\lfloor \cdot \rfloor$ is the floor operator element-wise. Let us remark that N takes the natural interpretation of the system's size. With this scaling, the number of servers increases directly proportionally to the number of jobs for each N, maintaining the relative ratios dictated by the initial densities in $\boldsymbol{\chi}(0)$. In order to keep up with the increasing service capacity, we define $\Gamma^i := \gamma^i N$ to be the arrival rate of the N-th Markov chain $(\mathbf{X}_N(t))_{t \geq 0}$. Then, the following result can be shown.

Theorem 1 *Fix arbitrary $T, \varepsilon > 0$, let $\mathcal{H}$ denote the set of all possible jumps of $(\mathbf{X}_N(t))_{t\geq 0}$ from (1) - (8) and let $\boldsymbol{\chi}(t)$ be the ODE solution of*

$$\frac{d}{dt}\boldsymbol{\chi}(t) = f(\boldsymbol{\chi}(t)) := \sum_{\mathbf{h}\in\mathcal{H}} \mathbf{h}\, f\big(\boldsymbol{\chi}(t), \mathbf{h}\big) \tag{10}$$

Then, it holds that $\lim_{N\to\infty} \mathbb{P}\big\{ \sup_{0\leq t\leq T} \big|\frac{1}{N}\mathbf{X}_N(t) - \boldsymbol{\chi}(t)\big| > \varepsilon\big\} = 0.$

Proof. Observing that for any $\mathbf{h} \in \mathcal{H}$ the function $\mathbf{x} \mapsto f(\mathbf{x}, \mathbf{h})$ is a sum of linear factors and minima or maxima thereof, it becomes apparent that $\mathbf{x} \mapsto f(\mathbf{x})$ is a globally Lipschitz continuous function. Hence, a direct application of Theorem 3.1 from [20] yields the claim. □

That is, the rescaled CTMCs converge in probability, as $N \to \infty$, to the ODE solution $\boldsymbol{\chi}$. The fluid solution can be used as an approximate to the average behavior of a CTMC with finite size N, e.g. the average number of jobs at station i is approximated by $N\big(\sum_{k=1}^{k_i}\sum_{l=1}^{l_i} C^i_{k,l}(t)\big)$. Note, however, that the above result ensures convergence only on finite time intervals, meaning that it does not necessarily imply the interchange of limits

$$\lim_{N\to\infty}\lim_{t\to\infty}\frac{1}{N}\mathbf{X}_N(t) = \lim_{t\to\infty}\lim_{N\to\infty}\frac{1}{N}\mathbf{X}_N(t)$$

In the next sections, we tackle this problem by identifying a sufficient condition that can be evaluated by means of efficient numerical algorithms for many queueing networks under study. Before doing so, however, we next derive the fluid models of our queuing networks. In particular, we separately discuss the fluid models of networks where service times or/and abandonment times are exponentially distributed. In the cases of exponential queuing networks, i.e. when $k_i = l_i = 1$ for all $1 \leq i \leq n$, the exposition greatly simplifies and the ideas behind the theory of LMIs become apparent. In particular, we will see that the presence of abandonments in the network is a necessary condition if one wants to use the LMI theory to prove that the stochastic steady state coincides with the global attractor of the ODE system in the limiting regime.

3.1 Networks with exponentially distributed service and abandonment times

In the case where $k_i = l_i = 1$ for all $1 \leq i \leq n$, it can be easily shown that the underlying fluid model is given by

$$\dot{C}^i = -\lambda^i C^i - \mu^i \min(C^i, S^i_1) + \sum_{j=1}^{n} r_{j,i}\mu^j \min(C^j, S^j_1) + \gamma^i, \qquad 1 \leq i \leq n \tag{11}$$

where we use the dot notation to indicate derivative with respect to time, and drop the explicit dependence on time in the ODEs.

Let us denote the drift (10) in the case of the above ODE system by f, with concentrations $S^1_1, \ldots, S^n_1$ and rates $\lambda^i, \mu^i, \gamma^i$, where $1 \leq i \leq n$, being fixed. Since f is piecewise affine, there exist affine functions $f_1, \ldots, f_k$ and a partition $\{\Delta_1, \ldots, \Delta_k\}$ of $\mathbb{R}^n$ such that $f(\mathbf{C}) = f_i(\mathbf{C})$ whenever $\mathbf{C} \in \Delta_i$. By expressing each affine function f_i in terms of the underlying matrix $A_i \in \mathbb{R}^{n\times n}$ and vector $b \in \mathbb{R}^n$, that is $f_i(\mathbf{C}) = A_i\mathbf{C} + b_i$ for all $\mathbf{C} \in \mathbb{R}^n$, it can be proven [33] that the ODE system $\dot{\mathbf{C}} = f(\mathbf{C})$ admits a global attractor

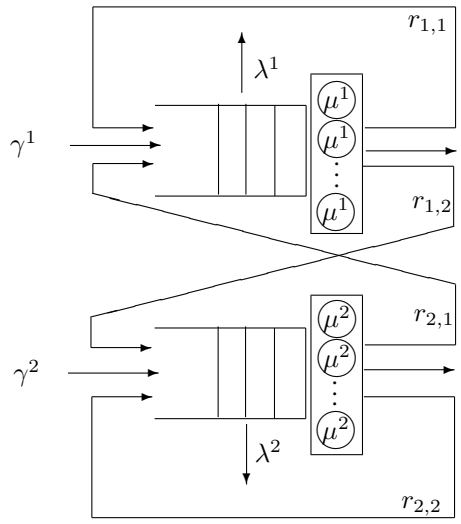


Fig. 1: Pictorial description of a tandem network with Poissonian arrivals and many-server queues that admit exponentially distributed service and abandonment times.

if there exists a symmetric positive definite matrix $P \in \mathbb{R}^{n\times n}$ which makes the matrices $PA_1 + A_1^T P, \ldots, PA_n + A_n^T P$ negative definite. That is, we look for a symmetric matrix $P > 0$ which solves the system of linear matrix inequalities (LMIs) $PA_i + A_i^T P < 0$, where $1 \leq i \leq n$. If the LMI system does not admit a solution, instead, it is unknown whether the ODE system has a global attractor or not.

Let us illustrate the procedure in the case of the tandem network given in Figure 1. Since the underlying ODE system is given by

$$\dot{C}^1 = -\lambda^1 C^1 - (1 - r_{1,1})\mu^1 \min(C^1, S_1^1) + r_{2,1}\mu^2 \min(C^2, S_1^2) + \gamma^1$$
$$\dot{C}^2 = -\lambda^2 C^2 + r_{1,2}\mu^1 \min(C^1, S_1^1) - (1 - r_{2,2})\mu^2 \min(C^2, S_1^2) + \gamma^2,$$

the corresponding sets Δ_i are given by

$$\Delta^1 = \{(C^1, C^2) \in \mathbb{R}^2 \mid C^1 \leq S_1^1 \wedge C^2 \leq S_1^2\}$$
$$\Delta^2 = \{(C^1, C^2) \in \mathbb{R}^2 \mid C^1 > S_1^1 \wedge C^2 > S_1^2\}$$
$$\Delta^3 = \{(C^1, C^2) \in \mathbb{R}^2 \mid C^1 \leq S_1^1 \wedge C^2 > S_1^2\}$$
$$\Delta^4 = \{(C^1, C^2) \in \mathbb{R}^2 \mid C^1 > S_1^1 \wedge C^2 \leq S_1^2\}$$

Consequently, the underlying matrices are

$$A_1 = \begin{pmatrix} -\lambda^1 - (1 - r_{1,1})\mu^1 & r_{2,1}\mu^2 \\ r_{1,2}\mu^1 & -\lambda^2 - (1 - r_{2,2})\mu^2 \end{pmatrix}, \quad A_2 = \begin{pmatrix} -\lambda^1 & 0 \\ 0 & -\lambda^2 \end{pmatrix},$$
$$A_3 = \begin{pmatrix} -\lambda^1 - (1 - r_{1,1})\mu^1 & 0 \\ r_{1,2}\mu^1 & -\lambda^2 \end{pmatrix}, \quad A_4 = \begin{pmatrix} -\lambda^1 & r_{2,1}\mu^2 \\ 0 & -\lambda^2 - (1 - r_{2,2})\mu^2 \end{pmatrix}$$

and the ODE system is guaranteed to have a global attractor if there exists a symmetric positive definite matrix $P \in \mathbb{R}^{2\times 2}$ such that $PA_i + A_i^T P < 0$ for all $1 \le i \le 4$. It can be shown that a *single* LMI $PA + A^T P < 0$ has a solution if and only if all eigenvalues of A have negative real parts [34, Section 1.2]. Consequently, if one of the matrices $A_1, \ldots, A_n$ is not invertible, the LMI system cannot have a solution. This, however, is the case for the LMI system of the tandem network given in Figure 1 if $\lambda^1 = 0$ or $\lambda^2 = 0$. Since this calculation carries over to networks of arbitrary size and topology, we conclude that queuing networks with feasible LMI systems need to have abandonment, an observation which is pivotal for the entire paper.

3.2 Networks with exponentially distributed service times and Coxian-distributed abandonment times

In the case where $k_i = 1$ for all $1 \le i \le n$, it can be easily shown that the underlying fluid model is given by

$$\begin{aligned}
\dot{C}_1^i &= -\lambda_1^i C_1^i - \mu^i \max\Big(\min\Big(C_1^i, S_1^i - \sum_{l'=2}^{l_i} C_{l'}^i\Big), 0\Big) \\
&\quad + \sum_{j=1}^{n} r_{j,i} \sum_{l=1}^{l_j} \mu^j \max\Big(\min\Big(C_l^j, S_1^j - \sum_{l'=l+1}^{l_j} C_{l'}^i\Big), 0\Big) + \gamma^i \\
\dot{C}_l^i &= -\lambda_l^i C_l^i - \mu^i \max\Big(\min\Big(C_l^i, S_1^i - \sum_{l'=l+1}^{l_i} C_{l'}^i\Big), 0\Big) + q_{l-1}^i \lambda_{l-1}^i C_{l-1}^i, \quad 2 \le l \le l_i
\end{aligned}$$

Together with $\tilde{l} := \min\{1 \le l \le l_i + 1 \mid \sum_{l'=l}^{l_i} C_{l'}^i \le S_1^i\}$, the service rate of the l-th phase satisfies

$$\mu^i \max\Big(\min\Big(C_l^i, S_1^i - \sum_{l'=l+1}^{l_i} C_{l'}^i\Big), 0\Big) = \begin{cases} 0 & , \ 1 \le l < \tilde{l} - 1 \\ \mu^i (S_1^i - \sum_{l'=\tilde{l}}^{l_i} C_{l'}^i) & , \ l = \tilde{l} - 1 \\ \mu^i C_l^i & , \ \tilde{l} \le l \le l_i \end{cases}$$

Consequently, queue i contributes $l_i + 1$ piecewise affine modes, meaning that the overall number of matrices underlying the drift is equal to $\prod_{i=1}^{n}(l_i + 1)$. Moreover, the dimension of each matrix is $(\sum_{i=1}^{n} l_i) \times (\sum_{i=1}^{n} l_i)$.

3.3 Networks with Coxian-distributed service and abandonment times

In order to simplify the exposition, we define $q_0^i := 0$ for all $1 \le i \le n$. Also, recall that $q_{l_i}^i := p_{k_i}^i := 0$. Then, it can be shown that the fluid model is given by the following ODE

system.

$$
\begin{aligned}
\dot{C}^i_{1,1} &= -\lambda^i_1 C^i_{1,1} - \mu^i_1 \max\Big(\min\Big(C^i_{1,1}, S^i_1 - \sum_{l'=2}^{l_i} C^i_{1,l'}\Big), 0\Big) \\
&\quad + \sum_{j=1}^{n} r_{j,i} \sum_{l=1}^{l_j} \Bigg[(1-p^j_1)\mu^j_1 \max\Big(\min\Big(C^j_{1,l}, S^j_1 - \sum_{l'=l+1}^{l_j} C^j_{1,l'}\Big), 0\Big) \\
&\quad + \sum_{k=2}^{k_j} (1-p^j_k)\mu^j_k C^j_{k,l}\Bigg] + \gamma^i, \\
\dot{C}^i_{1,l} &= -\lambda^i_l C^i_{1,l} - \mu^i_1 \max\Big(\min\Big(C^i_{1,l}, S^i_1 - \sum_{l'=l+1}^{l_i} C^i_{1,l'}\Big), 0\Big) + q^i_{l-1}\lambda^i_{l-1} C^i_{1,l-1}, \\
\dot{S}^i_1 &= -p^i_1 \sum_{l=1}^{l_i} \mu^i_1 \max\Big(\min\Big(C^i_{1,l}, S^i_1 - \sum_{l'=l+1}^{l_i} C^i_{1,l'}\Big), 0\Big) \\
&\quad + \sum_{l=1}^{l_i}\sum_{k=2}^{k_i} \Big((1-p^i_k)\mu^i_k C^i_{k,l} + (1-q^i_l)\lambda^i_l C^i_{k,l}\Big), \\
\dot{C}^i_{2,l} &= -\lambda^i_l C^i_{2,l} - \mu^i_2 C^i_{2,l} + p^i_1\mu^i_1 \max\Big(\min\Big(C^i_{1,l}, S^i_1 - \sum_{l'=l+1}^{l_i} C^i_{1,l'}\Big), 0\Big) \\
&\quad + q^i_{l-1}\lambda^i_{l-1} C^i_{2,l-1}, \\
\dot{C}^i_{k,l} &= -\lambda^i_l C^i_{k,l} - \mu^i_k C^i_{k,l} + p^i_{k-1}\mu^i_{k-1} C^i_{k-1} + q^i_{l-1}\lambda^i_{l-1} C^i_{k,l-1},
\end{aligned}
\tag{12}
$$

We make now the following important observation.

Remark 1 Note that (12) characterizes the fluid equilibria in full only if the initial condition is known. In particular, it holds that $\dot{S}^i_1 + \sum_{k=2}^{k_i}\sum_{l=1}^{l_i} \dot{C}^i_{k,l} = 0$ for all $1 \leq i \leq n$. However, by setting $S^i = S^i_1 + \sum_{k=2}^{k_i}\sum_{l=1}^{l_i} C^i_{k,l}$, it is possible to characterize the equilibria solely in terms of ODEs by removing the ODE of S^i_1 from (12) and by applying $S^i_1 = S^i - \sum_{k=2}^{k_i}\sum_{l=1}^{l_i} C^i_{k,l}$ to each occurrence of S^i_1, with $1 \leq i \leq n$.

It is exactly the modified ODE system from Remark 1 to which we will apply our result from the next section. That is, *after* fixing the server concentrations in each station and rewriting the ODE system as in Remark 1, we seek to establish the presence of a global attractor using the LMI theory. In the case this can be done, we infer that the system converges to the global attractor *regardless* of the initial concentration of the clients.

We end this section by discussing the complexity of the underlying fluid model. Similarly to the case where only the abandonment times were assumed to be Coxian-distributed, by setting

$$
\tilde{l} := \min\Big\{1 \leq l \leq l_i + 1 \mid \sum_{l'=l}^{l_i} C^i_{1,l'} \leq S^i - \sum_{k=2}^{k_i}\sum_{l=1}^{l_i} C^i_{k,l}\Big\}
$$

we observe that the service rate of service phase one of abandonment phase l-th satisfies

$$\mu_1^i \max\Big(\min\Big(C_{1,l}^i, S^i - \sum_{k=2}^{k_i}\sum_{l=1}^{l_i} C_{k,l}^i - \sum_{l'=l+1}^{l_i} C_{1,l'}^i\Big), 0\Big) =$$
$$= \begin{cases} 0 & , \ 1 \le l < \tilde{l} - 1 \\ \mu_1^i\Big(S^i - \sum_{k=2}^{k_i}\sum_{l=1}^{l_i} C_{k,l}^i - \sum_{l'=\tilde{l}}^{l_i} C_{1,l'}^i\Big) & , \ l = \tilde{l} - 1 \\ \mu_1^i C_{1,l}^i & , \ \tilde{l} \le l \le l_i \end{cases}$$

Thus, queue i contributes $l_i + 1$ piecewise affine modes as before, meaning that the size of the LMI system is equal to $\prod_{i=1}^n (l_i + 1)$. By making also the service times Coxian-distributed, however, the size of the ODE system and the matrices increase to $\sum_{i=1}^n l_i k_i$ and $(\sum_{i=1}^n l_i k_i) \times (\sum_{i=1}^n l_i k_i)$, respectively.

4 Analysis of the Steady State Regime

Under the assumption that the ODE system (12) admits a global attractor $\boldsymbol{\chi}^\star$, we next prove that the exists a sequence of steady-state measures $(\pi_N)_N$ underlying $(\mathbf{X}_N(t)/N)_N$ that converges, as $N \to \infty$, to $\boldsymbol{\chi}^\star$. Afterwards, we provide numerical evidence for the fact that the condition on global attraction holds true for a rich class of queuing networks.

Proof strategy. Due to the fact that the CTMCs $(\mathbf{X}_N(t)/N)_N$ have countable infinite state spaces, the common proof strategy [38, 16] does not apply. We address this by showing first in Proposition 1 that there exists a sequence of steady-state measures $(\pi_N)_N$ underlying $(\mathbf{X}_N(t)/N)_N$ by using results from the stability theory of Markov processes [31]. Building on that, we then prove in Theorem 2 that the aforementioned sequence is also tight. Armed with this, we extend the proof in [38, 16] and establish our main result, given as Theorem 3.

Proposition 1 *Let $(X(t))_{t\ge 0}$ denote an irreducible CTMC with a countable state space in $\mathbb{N}_0^d$ and $q_{v,w}$ be the transition rate from v into w, where we assume that each state v has a positive, finite number of successors w. Define the drift in v with respect to some measurable function $g : \mathbb{R}_{\ge 0}^d \to \mathbb{R}_{\ge 0}$ as $d_g(v) := \sum_{w:w\neq v}(g(w) - g(v))q_{v,w}$ and assume further that g satisfies the following conditions.*

1. *$g(v) \to \infty$ as $v \to \infty$*
2. *There exists a finite $K \subseteq \mathbb{N}_0^d$ and $c, d > 0$ such that $d_g(v) = \sum_{w:w\neq v}(g(w) - g(v))q_{v,w} \le -cg(v) + d$ for all v*
3. *It holds that $|d_g|(v) := \sum_{w:w\neq v} |g(w) - g(v)|q_{v,w} = \mathcal{O}(g(v))$*

Then the CTMC is positive recurrent and the unique steady-state measure π is such that $\mathbb{E}_\pi[g(X(0))] < \infty$. Moreover, it holds that $0 = \sum_v \pi(v) d_g(v)$.

Proof of Proposition 1. The ergodicity of the CTMC and the fact that $\mathbb{E}_\pi[g(X(0))] < \infty$ are due to Theorem 7.1 of [31]. Moreover, Theorem 2 of [15] ensures that

$$\mathbb{E}_\alpha[g(X(t))] = \mathbb{E}_\alpha[g(X(0))] + \mathbb{E}_\alpha\left[\int_0^t d_g(X(s))ds\right] = \mathbb{E}_\alpha[X(0)] + \int_0^t \mathbb{E}_\alpha[d_g(X(s))]ds \tag{13}$$

if $\mathbb{E}_\alpha[g(X(0))] < \infty$, meaning that $\mathbb{E}_\alpha[|d_g|(X(t))] = \sum_v \mathbb{P}_\alpha(X(t) = v)|d_g|(v) \leq c\sum_v \mathbb{P}_\alpha(X(t) = v)g(v) = c\mathbb{E}_\alpha[g(X(t))]$ for some $c > 0$. Thus, (13) and Gronwall's inequality ensure the existence of some $C > 0$ such that $\mathbb{E}_\alpha[g(X(t))] \leq \mathbb{E}_\alpha[g(X(0))] + C(e^{ct} - 1)$ for all $t \geq 0$, which readily implies that $s \mapsto \mathbb{E}_\alpha[d_g(X(s))]$ is integrable on any compact interval. Hence, the fundamental theorem of Lebesgue calculus and (13) imply that $\dot{\mathbb{E}}_\alpha[g(X)] = \mathbb{E}_\alpha[d_g(X)]$ almost everywhere on $\mathbb{R}_{\geq 0}$. The last claim then follows from $\dot{\mathbb{E}}_\pi[g(X(t))] = \mathbb{E}_\pi[d_g(X(t))] = \sum_v \mathbb{P}_\pi(X(t) = v)d_g(v) = \sum_v \pi(v)d_g(v)$. □

Using Proposition 1, we can prove that our networks induce a tight sequence of measures.

Theorem 2 *For each $N \geq 1$, the N-th queuing network $(X_N(t)/N)_{t\geq 0}$ has a steady-state measure π_N. Moreover, the sequence $(\pi_N)_N$ is tight, meaning that for any $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subseteq \mathbb{R}^{\sum_{i=1}^n l_i k_i}$ such that $\pi_N(K_\varepsilon) \geq 1 - \varepsilon$ for all $N \geq 1$.*

Proof of Theorem 2. Recall that the populations $X_{S_1^i}$ can be removed from the state descriptor $\mathbf{X}$ by fixing an initial server population X_{S^i} thanks to the relation $X_{S^i} = X_{S_1^i} + \sum_{k=2}^{k_i}\sum_{l=1}^{l_i} X_{C_{k,l}^i}$. Thus, one can assume that $\mathbf{X} = (X_{C_{k,l}^i})_{1\leq i\leq n, 1\leq l\leq l_i, 1\leq k\leq k_i}$. In the following, $\mathbf{Z}$ and $\mathbf{Z}'$ denote states of $(\mathbf{X}_N(t)/N)_{t\geq 0}$, whereas $g_{C_{k,l}^i}(\mathbf{Z}) := Z_{C_{k,l}^i}$ and $g := \sum_{C_{k,l}^i} g_{C_{k,l}^i}$. Further, let $F_{C_{k,l}^i}$ be the ODE formula which refers to the change of $X_{C_{k,l}^i}(t)/N$ in time, e.g.

$$F_{C_{1,l}^i}(\mathbf{Z}) = -\mu_1^i \max\Big(\min\Big(Z_{C_{1,l}^i}, Z_{S^i} - \sum_{k=2}^{k_i}\sum_{l=1}^{l_i} Z_{C_{k,l}^i} - \sum_{l'=l+1}^{l_i} Z_{C_{1,l'}^i}\Big), 0\Big) - \lambda_l^i Z_{C_{1,l}^i} + q_{l-1}^i \lambda_{l-1}^i Z_{C_{1,l-1}^i}$$

Note that the fluid approximation is given by the equation

$$\sum_{\mathbf{Z}':\mathbf{Z}'\neq\mathbf{Z}} (g_{C_{k,l}^i}(\mathbf{Z}') - g_{C_{k,l}^i}(\mathbf{Z}))q_{\mathbf{Z},\mathbf{Z}'} = F_{C_{k,l}^i}(\mathbf{Z})$$

Similarly, it can be shown that $\sum_{\mathbf{Z}':\mathbf{Z}'\neq\mathbf{Z}} |g_{C_{k,l}^i}(\mathbf{Z}') - g_{C_{k,l}^i}(\mathbf{Z})|q_{\mathbf{Z},\mathbf{Z}'}$ arises from $F_{C_{k,l}^i}(\mathbf{Z})$ by turning any minus into a plus. Note that the latter implies the third condition of Proposition 1. The former, instead, induces

$$\begin{aligned}\sum_{C_{k,l}^i} F_{C_{k,l}^i}(\mathbf{Z}) &= \sum_{C_{k,l}^i}\sum_{\mathbf{Z}':\mathbf{Z}'\neq\mathbf{Z}} (g_{C_{k,l}^i}(\mathbf{Z}') - g_{C_{k,l}^i}(\mathbf{Z}))q_{\mathbf{Z},\mathbf{Z}'} \\ &= \sum_{\mathbf{Z}':\mathbf{Z}'\neq\mathbf{Z}}\sum_{C_{k,l}^i} (g_{C_{k,l}^i}(\mathbf{Z}') - g_{C_{k,l}^i}(\mathbf{Z}))q_{\mathbf{Z},\mathbf{Z}'} \\ &= \sum_{\mathbf{Z}':\mathbf{Z}'\neq\mathbf{Z}} (g(\mathbf{Z}') - g(\mathbf{Z}))q_{\mathbf{Z},\mathbf{Z}'} = d_g(\mathbf{Z}).\end{aligned}$$

Consequently, it suffices to find some $c, d > 0$ such that $\sum_{C_{k,l}^i} F_{C_{k,l}^i}(\mathbf{Z}) \leq c\sum_{C_{k,l}^i} Z_{C_{k,l}^i} + d$ for all $\mathbf{Z}$ in order to infer the second condition of Proposition 1. For this, we observe that

$$\sum_{C_{k,l}^i} F_{C_{k,l}^i}(\mathbf{Z}) \leq -(1-\overline{q})\underline{\lambda}\sum_{C_{k,l}^i} Z_{C_{k,l}^i} + \gamma_\Sigma = -(1-\overline{q})\underline{\lambda}\cdot g(\mathbf{Z}) + \gamma_\Sigma$$

where $\overline{q} := \max\{q_l^i \mid 1 \le i \le n \wedge 1 \le l \le l_i\}$, $\underline{\lambda} := \min\{\lambda_l^i \mid 1 \le i \le n \wedge 1 \le l \le l_i\}$ and $\gamma_\Sigma := \sum_{i=1}^n \gamma^i$. Since g satisfies also the first condition of Proposition 1, we can apply the latter to infer the existence of π_N. To see the second part of the claim, we define $K_\varepsilon := [0; 1/\varepsilon]^{\sum_{i=1}^n l_i k_i}$ and observe that for any $\eta > 0$ there exists a sufficiently small $\varepsilon > 0$ such that $d_g(\mathbf{Z}) \le -(1-\overline{q})\underline{\lambda}\cdot g(\mathbf{Z}) + \gamma_\Sigma = -(1-\overline{q})\underline{\lambda}\cdot\|\mathbf{Z}\|_1 + \gamma_\Sigma \le -\eta + (\eta+\gamma_\Sigma)\cdot \mathbb{1}_{K_\varepsilon}(Z)$ for all $\mathbf{Z}$. This yields $\sum_{\mathbf{Z}} \pi_N(\mathbf{Z}) d_g(\mathbf{Z}) \le -\eta + (\eta+\gamma_\Sigma)\pi_N(K_\varepsilon)$. Thanks to Proposition 1, it holds that $\sum_{\mathbf{Z}} \pi_N(\mathbf{Z}) d_g(\mathbf{Z}) = 0$ and we can use the idea from [11] to infer $\frac{\eta}{\eta+\gamma_\Sigma} \le \pi_N(K_\varepsilon)$. Since the choice of ε does not depend on N, the proof is complete. □

Armed with Theorem 2, we are able to extend the proof strategy used in [38, 16] to the case of our networks.

Theorem 3 *Let us assume that the ODE system underlying the network family $(\frac{1}{N}\mathbf{X}_N(t))_{t\ge 0}$ has a unique global attractor $\boldsymbol{\chi}^\star$. Then, for any $\delta > 0$, it holds that*

$$\lim_{N\to\infty} \lim_{t\to\infty} \mathbb{P}\Big\{ \Big| \frac{1}{N}\mathbf{X}_N(t) - \boldsymbol{\chi}^\star \Big| > \delta \Big\} = 0$$

Proof. The following proof is a modification of the proofs given in [38, 16] which, in turn, are based on [4]. Let $\boldsymbol{\chi}_{\mathbf{x}}(t)$ denote the ODE solution subject to $\boldsymbol{\chi}_{\mathbf{x}}(0) = \mathbf{x}$. Further, for any $A \subseteq \mathbb{R}_{\ge 0}^{\sum_i k_i l_i}$, we define $\boldsymbol{\chi}_A^{-1}(t) := \{\mathbf{x} \in \mathbb{R}_{\ge 0}^{\sum_i k_i l_i} \mid \boldsymbol{\chi}_{\mathbf{x}}(t) \in A\}$. Thanks to the Theorem 2, we know that there exists a sequence of steady-state measures $(\pi_N)_N$ underlying $(\frac{1}{N}\mathbf{X}_N(t))_{t\ge 0}$ that is tight. Hence, we can use Prokhorov's theorem to fix a subsequence $(\pi_{N_i})_i$ of $(\pi_N)_N$ which converges weakly against some measure π. Next, we show that π is a steady-state measure for $\boldsymbol{\chi}(t)$, that is $\pi\big(\boldsymbol{\chi}_A^{-1}(t)\big) = \pi(A)$ for any Borel-measurable set A of $\mathbb{R}_{\ge 0}^{\sum_i k_i l_i}$ and $t \ge 0$. Obviously, $\pi\big(\boldsymbol{\chi}_A^{-1}(t)\big) = \pi(A)$ is equivalent to

$$\int_{\mathbb{R}_{\ge 0}^{\sum_i k_i l_i}} \mathbb{1}_A(\boldsymbol{\chi}_{\mathbf{x}}(t))\pi(d\mathbf{x}) = \int_{\mathbb{R}_{\ge 0}^{\sum_i k_i l_i}} \mathbb{1}_A(x)\pi(d\mathbf{x}) \tag{14}$$

To show the above equation, one first proves that

$$\int_{\mathbb{R}_{\ge 0}^{\sum_i k_i l_i}} g(\boldsymbol{\chi}_{\mathbf{x}}(t))\pi(d\mathbf{x}) = \int_{\mathbb{R}_{\ge 0}^{\sum_i k_i l_i}} g(x)\pi(d\mathbf{x}) \tag{15}$$

holds true for any bounded uniformly continuous function $g : \mathbb{R}_{\ge 0}^{\sum_i k_i l_i} \to \mathbb{R}$. The corresponding proof can be taken verbatim from [16]. Having this, we observe that, for any $\varepsilon > 0$, $g_\varepsilon(\cdot) := \min\big(1, d(\cdot, A)/\varepsilon\big)$ is a bounded uniformly continuous function, where $d(\mathbf{x}, A) := \inf\{d(\mathbf{x}, \mathbf{a}) \mid \mathbf{a} \in A\}$ denotes the Euclidian distance between $\mathbf{x}$ and A. This holds true because $|d(\mathbf{x}, A) - d(\mathbf{y}, A)| \le d(\mathbf{x}, \mathbf{y})$. (To see this, note that $d(\mathbf{x}, \mathbf{a}) \le d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{a})$ for any $\mathbf{a} \in A$; taking infimum over $\mathbf{a} \in A$ yields then the inequality.) Thus, by applying the theorem of dominated convergence, we infer that $\int g_\varepsilon(\mathbf{x})\pi(d\mathbf{x}) \to \int \mathbb{1}_A(\mathbf{x})\pi(d\mathbf{x})$ as $\varepsilon \to 0$. This and (15) implies the desired equation (14). Let us denote in the sequel the closed ball around $\boldsymbol{\chi}^\star$ with radius $\varepsilon > 0$ by $B_\varepsilon(\boldsymbol{\chi}^\star)$. Thanks to the fact that $\boldsymbol{\chi}^\star$ is a unique global attractor, for any $\varepsilon > 0$ and any compact set $K \subseteq \mathbb{R}_{\ge 0}^{\sum_i k_i l_i}$ there exists a $t > 0$ such that $\boldsymbol{\chi}_K(t) := \{\boldsymbol{\chi}_{\mathbf{x}}(t) \mid \mathbf{x} \in K\} \subseteq B_\varepsilon(\boldsymbol{\chi}^\star)$. Since this implies $K \subseteq \chi^{-1}_{B_\varepsilon(\boldsymbol{\chi}^\star)}(t)$ and, by Theorem 1.4 from [5], there exists for any $\eta > 0$ a compact set K_η such that $\pi(K_\eta) \ge 1 - \eta$, we infer that for any $\varepsilon, \eta > 0$ there exists a $t > 0$ with $\pi\big(\chi^{-1}_{B_\varepsilon(\boldsymbol{\chi}^\star)}(t)\big) \ge 1 - \eta$. Noting that this and (14) imply $\pi\big(B_\varepsilon(\boldsymbol{\chi}^\star)\big) \ge 1 - \eta$ for all $\varepsilon, \eta > 0$, this yields $\pi = \delta_{\boldsymbol{\chi}^\star}$, i.e. $(\pi_N)_N$ converges weakly to the Dirac measure $\delta_{\boldsymbol{\chi}^\star}$. This and the proof of Theorem 5 in [38] yield then the claim. □

Note that if $\dot{\boldsymbol{\chi}} = F(\boldsymbol{\chi})$ admits a global attractor, then $F(\boldsymbol{\chi}) = 0$ has a unique solution, namely the global attractor itself. Although the converse is not true in general (consider, for instance, the single ODE $\dot{x} = x$), it is interesting to ask whether our ODE system is such that $F(\boldsymbol{\chi}) = 0$ has always a unique solution. We study this question in the case of exponentially distributed service and abandonment times. The next proposition states that there exists always a solution; uniqueness, instead, is shown under an additional assumption.

Proposition 2 *Let $\dot{\boldsymbol{\chi}} = F(\boldsymbol{\chi})$ denote the ODE system of (11). Then, the system of equations $F(\boldsymbol{\chi}) = 0$ admits a solution. Moreover, if $\min_i r_{i,i} = p$, then $F(\boldsymbol{\chi}) = 0$ admits a unique solution if $2(1-p)\max_i \mu^i < \min_i \lambda^i$.*

Proof. Together with $R^T = (r_{j,i})_{i,j}$, $\mathbf{T_C} = (\mu^1 \min(|C^1|, S^1), \ldots, \mu^n \min(|C^n|, S^n))$, $\lambda^{-1}(\mathbf{X}) = (X^1/\lambda^1, \ldots, X^n/\lambda^n)$ and $\boldsymbol{\gamma} = (\gamma^1, \ldots \gamma^n)$, the equation $F(\mathbf{C}) = 0$ rewrites to $\lambda^{-1}((R^T - I)\mathbf{T_C} + \boldsymbol{\gamma}) = \mathbf{C}$. Let us consider the function

$$\Theta : \mathbb{R}^n \to \mathbb{R}^n, \quad \mathbf{C} \mapsto \lambda^{-1}\big((R^T - I)\mathbf{T_C} + \boldsymbol{\gamma}\big)$$

Thanks to Schauder's fixed point theorem, Θ has a fixed point, meaning that $F(\boldsymbol{\chi}) = 0$ admits a solution. Further, it holds that

$$\begin{aligned}
\|\Theta(\mathbf{C}) - \Theta(\mathbf{C}')\|_1 &= \|\lambda^{-1}\big((R^T - I)(\mathbf{T_C} - \mathbf{T_{C'}})\big)\|_1 \\
&= \|\lambda^{-1}\big((R^T - I)\big)(\mathbf{T_C} - \mathbf{T_{C'}})\|_1 \\
&\leq \|\lambda^{-1}\big((R^T - I)\big)\|_1 \|\mathbf{T_C} - \mathbf{T_{C'}}\|_1 \\
&\leq \frac{\max_i \mu^i}{\min_i \lambda^i} \|R^T - I\|_1 \|\mathbf{C} - \mathbf{C}'\|_1,
\end{aligned}$$

meaning that Banach's fixed point theorem ensures the existence of a unique equilibrium if the assumption is fulfilled. □

For instance, if a customer is redirected to the same queue after being served in at least 75% of all cases, then the condition in Proposition 2 rewrites to $\max_i \mu^i < 2 \min_i \lambda^i$ which says, essentially, that the average patience of a customer lasts twice as long as the average service time.

Note that the assumptions made in Proposition 2 are used to invoke Banach's fixed point theorem, thus they do not provide one with further insights to the model but are merely a sufficient condition for the proof of the proposition to go through. At the same time, however, all models that have been considered by the authors seemed to have unique equilibrium points. Although it would be interesting to extend Proposition 2 to networks with Coxian service or abandonment times, the proof seems not to generalize.

5 Numerical Assessment

In this section we provide numerical evidence that the existence and uniqueness of a global attractor can be successfully established through an LMI feasibility problem.

Before doing so, however, we first demonstrate our approach on the queueing network depicted in Figure 2. Its 2-Coxian services are given by the rates $(\mu_1^1, \mu_2^1) = (2.0, 0.2)$, $(\mu_1^2, \mu_2^2) = (1.0, 0.2)$, $(\mu_1^3, \mu_2^3) = (0.5, 0.2)$, $(\mu_1^4, \mu_2^4) = (1.0, 0.2)$, $(\mu_1^5, \mu_2^5) = (0.5, 0.2)$ and the probabilities $p_1^1 = 0.1$, $p_1^2 = 0.2$, $p_1^3 = 0.3$, $p_1^4 = 0.2$, $p_1^5 = 0.3$. Instead, the 2-Coxian abandonments are given by $(\lambda_1^i, \lambda_2^i) = (0.2, 0.2)$ and $q_1^i = 1.0$, where $1 \leq$

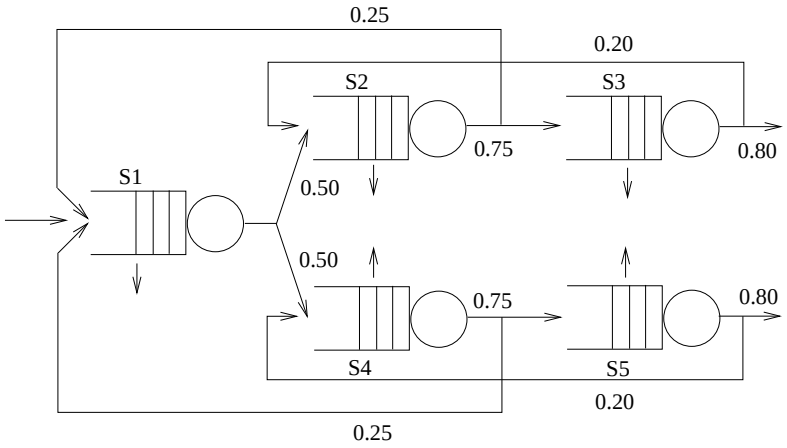
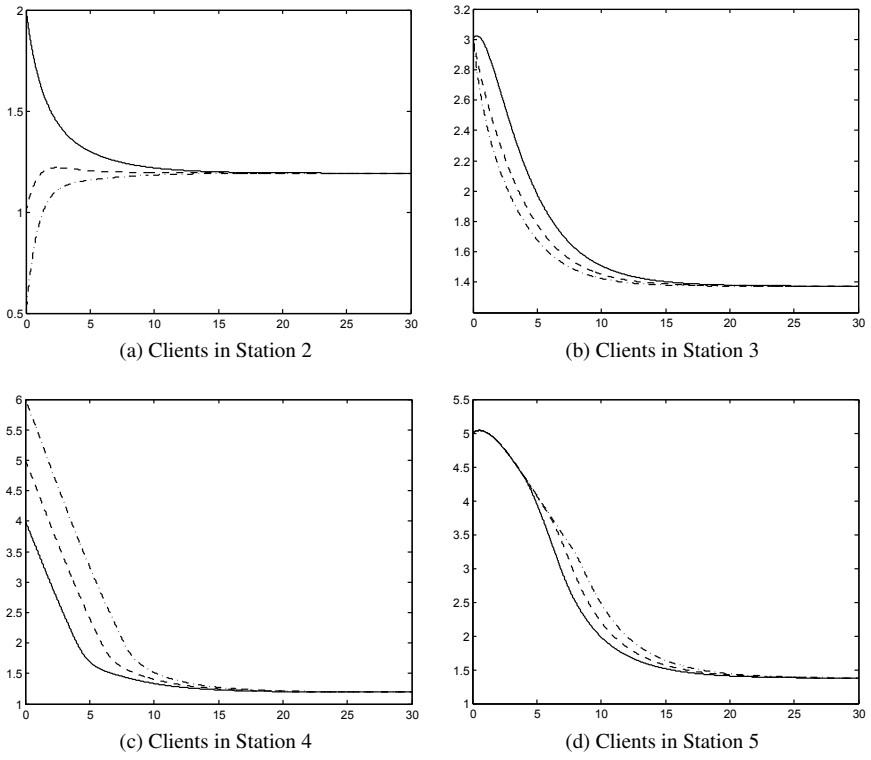


Fig. 2: Case study queueing network with 2-Coxian service and abandonment times.



(a) Clients in Station 2

(b) Clients in Station 3

(c) Clients in Station 4

(d) Clients in Station 5

Fig. 3: ODE solutions underlying the network from Figure 2. Solid lines, dashed lines and dash-dot lines refer to initial conditions a), b) and c), respectively. The trajectories of all three initial conditions converge to a common attractor.

	Exp / Exp		2-Cox / Exp		Exp / 2-Cox		2-Cox / 2-Cox	
# Queues	*Feas.*	*Time*	*Feas.*	*Time*	*Feas.*	*Time*	*Feas.*	*Time*
3	99%	0s	98%	0s	91%	0s	83%	0s
4	99%	0s	98%	0s	92%	0s	69%	2s
5	99%	0s	98%	0s	66%	1s	44%	31s
6	99%	0s	98%	0s	48%	10s	-	-
7	99%	0s	96%	1s	41%	75s	-	-
8	98%	0s	89%	5s	-	-	-	-
9	89%	1s	83%	24s	-	-	-	-
10	88%	4s	81%	75s	-	-	-	-
11	88%	16s	-	-	-	-	-	-

Table 1: Percentage of queuing networks for which global attraction could be established through an LMI feasibility problem. The header row shows which combination of service- and abandonment-time distribution was considered.

$i \leq 5$. The arrival rate to station one was set to 5.0, while the initial concentrations of the servers were chosen to be $S_1^1(0) = 1.0$, $S_1^2(0) = 2.0$, $S_1^3(0) = 3.0$, $S_1^4(0) = 2.0$, $S_1^5(0) = 3.0$. We remark that if Y_i^S and Y_i^A denote the random variables describing the service and abandonment times in station $1 \leq i \leq 5$, respectively, the above parameters yield $\mathbb{E}[Y_1^S] = 1$, $\mathbb{V}[Y_1^S] = 5$, $\mathbb{E}[Y_2^S] = 2$, $\mathbb{V}[Y_2^S] = 10$, $\mathbb{E}[Y_3^S] = 4$, $\mathbb{V}[Y_3^S] = 20$, $\mathbb{E}[Y_4^S] = 2$, $\mathbb{V}[Y_4^S] = 10$, $\mathbb{E}[Y_5^S] = 4$, $\mathbb{V}[Y_5^S] = 20$ and $\mathbb{E}[Y_i^A] = 10$, $\mathbb{V}[Y_i^A] = 50$, as can be easily shown by invoking the Laplace transform [35, Section 7.6.7].

The fluid model (12) of our network has 25 ODEs. However, as discussed in Remark 1, by fixing the initial server concentrations, the ODE system can be rewritten into one of size 20 by eliminating the ODE variables $S_1^1, \ldots, S_1^5$. The remaining ODE variables are thus those of the clients $(C_{1,1}^i, C_{1,2}^i, C_{2,1}^i, C_{2,2}^i)_{1 \leq i \leq 5}$. Figure 3 depicts the total client concentration $C_{1,1}^i + C_{1,2}^i + C_{2,1}^i + C_{2,2}^i$ present in station i, where $2 \leq i \leq 5$, for the following three initial conditions:

a) $C_1^1(0) = 1.0$, $C_1^2(0) = 2.0$, $C_1^3(0) = 3.0$, $C_1^4(0) = 4.0$, $C_1^5(0) = 5.0$ and 0 otherwise.
b) $C_1^1(0) = 1.0$, $C_1^2(0) = 1.0$, $C_1^3(0) = 3.0$, $C_1^4(0) = 5.0$, $C_1^5(0) = 5.0$ and 0 otherwise.
c) $C_1^1(0) = 1.0$, $C_1^2(0) = 0.5$, $C_1^3(0) = 3.0$, $C_1^4(0) = 6.0$, $C_1^5(0) = 5.0$ and 0 otherwise.

The trajectories in Figure 3 suggest that the ODE system admits a global attractor. Indeed, using the Robust Control Toolbox version 5.2 of Matlab version R2014b, it is possible to show that the LMI system underlying the ODE system is feasible, i.e. has a solution. (The underlying LMI solution is in $\mathbb{R}^{20 \times 20}$ and is omitted due to space reasons.) This ensures that the attractor found by solving the ODE system for a), b) or c) is a global attractor and that *any* initial client concentration will converge to it. More importantly, Theorem 3 allows us to conclude that the sequence of steady-state measures converges to the global attractor.

Since Theorem 3 can be applied only in the presence of a global attractor, we next provide numerical evidence that the existence and uniqueness of a global attractor can be successfully established through an LMI feasibility problem. For this, we constructed LMI systems of 100 queuing networks with randomly chosen parameters for a fixed network size, which we varied from 3 to 11. For each network we randomly generated the mean service and abandonment times such that $1/\mathbb{E}[Y_i^S] \sim \mathcal{U}(1.0; 3.0)$ and $1/\mathbb{E}[Y_i^A] \sim \mathcal{U}(0.001; 2.0)$,

where $\mathcal{U}(a;b)$ denotes the uniform distribution on interval $(a;b)$. This means that the average patience of a customer lasted twice as long as the average service time. Then, we considered the four possible combinations of the service time distributions for service and abandonments, obtained by choosing from an exponential and a two-stage Coxian, with coefficient of variation drawn from the uniform distribution $\mathcal{U}(0.5;10.0)$. The routing probability mass $r^i \sim \mathcal{U}[0.7;1.0]$ was spread randomly across $r_{i,1},\ldots,r_{i,n}$. Here, we would like to point out that two-stage Coxians allow to approximate a rich class of distributions. In particular, if $E,V>0$ are such that $V/E^2 \geq 0.5$, a two-stage Coxian random variable Y can be constructed [35, Section 7.6.7] such that $\mathbb{E}[Y]=E$ and $\mathbb{V}[Y]=V$.

Our findings are summarized in Table 1, grouped in columns according to the combination of service- and abandonment-time distribution considered, where each rows shows the statistics for a given network size. The table lists the percentage of networks which induce a feasible LMI system (and hence proving the presence of a global attractor) and the average time to run the analysis for a single network. This was measured on an ordinary laptop equipped with an Intel Core i5-3210M processor and 8 GB RAM. Entries in the table indicated by '-' refer to cases where either an out-of-memory error was produced, or where the analysis of a single network exceeded 180 s, an arbitrarily chosen time bound.

These results confirm the increased computational cost of the analysis due to the matrix sizes and/or the number of linear modes in the ODEs, as discussed in Sections 3.1–3.3. For a fixed combination of service- and abandonment-time distribution, increasing the number of ODE linear modes leads to a decrease in the percentage of feasible LMI problems. We explain this phenomenon with the growing number of LMI equations that have to admit a common solution.

6 Mixed Abandonment Policy

In this section we extend our exponential model (11) and allow the abandonment distribution to depend on the fact whether a customer is waiting or being served. While this usually does not apply to computer systems where clients are jobs and abandonment is timeout, in the case of call centers, this can be explained by the changed mindset of a customer that entered service. We think that this policy is more flexible than the common abandonment policy [28] according to which a customer cannot abandon while being served.

In particular, instead of considering the fluid model (11) where the abandonment rate while waiting and while being served in station i is in both cases λ^i, we study the situation where the former is equal to λ^i_w and the latter is given by λ^i_s. It can be easily seen that the abandonment policy of [28] can be extended to such a case and gives rise to the ODE system

$$\begin{aligned}\dot{C}^i = &-\mu^i \min(C^i, S^i_1) + \sum_{j=1}^{n} r_{j,i}\mu^j \min(C^j, S^j_1) \\ &- \lambda^i_w \max(C^i - S^i_1, 0) - \lambda^i_s \min(C^i, S^i_1) + \gamma^i, \end{aligned} \tag{16}$$

where $1 \leq i \leq n$. Exactly as in Section 3.1, we first ask ourselves whether the LMI system underlying (16) can be feasible. With $\Delta_1,\ldots,\Delta_4$ and $A_1,\ldots,A_4$ as in Section 3.1, the LMI

	$\mu^i \sim \mathcal{U}(1.0;3.0)$ $\lambda_w^i \sim \mathcal{U}(0.001;2.0)$ $\lambda_s^i = 0$		$\mu^i \sim \mathcal{U}(1.0;3.0)$ $\lambda_w^i \sim \mathcal{U}(0.001;2.0)$ $\lambda_s^i \sim \mathcal{U}(0.0001;0.002)$	
# Queues	*Feas.*	*Time*	*Feas.*	*Time*
3	97%	0s	97%	0s
4	96%	0s	97%	0s
5	95%	0s	96%	0s
6	94%	0s	95%	0s
7	93%	0s	96%	0s
8	93%	1s	93%	0s
9	67%	2s	85%	2s
10	64%	9s	79%	5s
11	66%	27s	79%	20s

Table 2: Percentage of queuing networks for which global attraction could be established through an LMI feasibility problem. The header row shows which combination of service- and abandonment-rates has been studied.

system of (16) for $n = 2$ is

$$A_1 = \begin{pmatrix} -\lambda_s^1 - (1-r_{1,1})\mu^1 & r_{2,1}\mu^2 \\ r_{1,2}\mu^1 & -\lambda_s^2 - (1-r_{2,2})\mu^2 \end{pmatrix}, \quad A_2 = \begin{pmatrix} -\lambda_w^1 & 0 \\ 0 & -\lambda_w^2 \end{pmatrix},$$
$$A_3 = \begin{pmatrix} -\lambda_s^1 - (1-r_{1,1})\mu^1 & 0 \\ r_{1,2}\mu^1 & -\lambda_w^2 \end{pmatrix}, \quad A_4 = \begin{pmatrix} -\lambda_w^1 & r_{2,1}\mu^2 \\ 0 & -\lambda_s^2 - (1-r_{2,2})\mu^2 \end{pmatrix}$$

As in Section 3.1, we note that $\lambda_w^1, \lambda_w^2 > 0$ is necessary for the feasibility of the LMI system. Since this means that waiting customers should be able to abandon, we infer that the common abandonment policy satisfies the necessary condition. Note, however, that the common abandonment policy implies also that $\lambda_s^1 = \lambda_s^2 = 0$. Consequently, if the service rates μ^1, μ^2 are too small, the LMI system could become infeasible.

Following a similar route as in Section 5, we investigated the relation between the likelihood of feasibility and the mean abandonment times during service. To this end, we generated for each network size $3 \leq n \leq 11$ randomly 100 queueing networks and calculated the average computation time and the underlying percentage of feasibility. The routing probabilities were generated as in Section 5. The results are depicted in Table 2 and allow to draw three important conclusions. First, the LMI approach applies well to the common abandonment policy where customers cannot abandon while being served. This can be observed by studying the first column of Table 2. Second, already very large mean abandonment times in service substantially improve the chances of feasibility. This is underpinned by column two of Table 2. Third, by decreasing the mean abandonment times in service, the feasibility can be further improved. This can be seen by comparing the second column of Table 2 with the first column of Table 1, whose rates were drawn according to $\mu^i \sim \mathcal{U}(1.0;3.0)$ and $\lambda_w^i, \lambda_s^i \sim \mathcal{U}(0.001;2.0)$.

7 Conclusion

In this paper we proposed a computational method to study the convergence in the steady state to the fluid limit in the sense of Kurtz for queuing networks with Coxian-distributed service and abandonment times. This is made possible by the fact that the limit ODE system has a piecewise affine vector field. Thus, the existence of a unique global attractor, which is a sufficient condition for convergence in the steady state, can be established by setting up a feasibility problem for LMIs. An empirical evaluation on a collection of queuing networks with randomly generated parameters has shown its applicability in practice. In particular, our analysis has highlighted that the presence of abandonment is necessary for feasibility. Future work will aim at producing a tool implementation to support the findings herein presented. Moreover, we want to investigate the possibility of extending mixed abandonment times to non-exponential queueing networks.

Acknowledgement This work was partially supported by the EU project QUANTICOL, 600708.

References

1. Anselmi, J., Verloop, I.: Energy-aware capacity scaling in virtualized environments with performance guarantees. Performance Evaluation **68**(11), 1207–1221 (2011)
2. Ascher, U.M., Petzold, L.R.: Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations. SIAM (1988)
3. Baskett, F., Chandy, K.M., Muntz, R.R., Palacios, F.G.: Open, Closed, and Mixed Networks of Queues with Different Classes of Customers. J. ACM **22**(2), 248–260 (1975)
4. Benaim, M.: Recursive algorithms, urn processes and chaining number of chain recurrent sets. Ergodic Theory and Dynamical Systems **18**, 53–87 (1998)
5. Billingsley, P.: Convergence of probability measures, second edn. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc. (1999). A Wiley-Interscience Publication
6. Bolch, G., Greiner, S., de Meer, H., Trivedi, K.: Queueing networks and Markov chains: modeling and performance evaluation with computer science applications. Wiley (2005)
7. Chen, H., Yao, D.D.: Fundamentals of Queueing Networks. Springer (2001)
8. Cumani, A.: On the canonical representation of homogeneous Markov processes modelling failure-time distributions. Microelectronics Reliability **22**(3), 583–602 (1982)
9. Dai, J., Dieker, A., Gao, X.: Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. Queueing Systems **78**(1), 1–29 (2014). DOI 10.1007/s11134-014-9394-x
10. Dai, J., He, S.: Many-server queues with customer abandonment: A survey of diffusion and fluid approximations. Journal of Systems Science and Systems Engineering **21**(1), 1–36 (2012)
11. Dayar, T., Hermanns, H., Spieler, D., Wolf, V.: Bounding the equilibrium distribution of Markov population models. Numerical Linear Algebra with Applications **18**(6), 931–946 (2011)
12. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: Tutorial, review, and research prospects. Manufacturing & Service Operations Management **5**(2), 79–141 (2003)
13. Gast, N., Gaujal, B.: A Mean Field Model of Work Stealing in Large-Scale Systems. In: SIGMETRICS, pp. 13–24 (2010)
14. Halfin S. and Whitt W.: Heavy-traffic limits theorem for queues with many exponential servers. Operations Research **29**, pp. 567–588 (1981)
15. Hamza, K., Klebaner, F.C.: Conditions for Integrability of Markov Chains. Journal of Applied Probability **32**(2), pp. 541–547 (1995)
16. Hayden, R.: Convergence of ODE approximations and bounds on performance models in the steady-state. In: Ninth Workshop on Process Algebra and Stochastically Timed Activities (PASTA) (2010)
17. Jennings, O.B., Puha, A.L.: Fluid limits for overloaded multiclass FIFO single-server queues with general abandonment. Stochastic Systems **3**(1), 262–321 (2013)
18. Kang, W., Pang, G.: Fluid Limit of A Many-Server Queueing Network with Abandonment (2013). Submitted
19. Kang, W., Ramanan, K.: Fluid limits of many-server queues with reneging. The Annals of Applied Probability **200**(6), 2204–2260 (2010)

20. Kurtz, T.G.: Solutions of ordinary differential equations as limits of pure Markov processes. J. Appl. Prob. **7**(1), 49–58 (1970)
21. Liu, Y., Whitt, W.: A Network of Time-Varying Many-Server Fluid Queues with Customer Abandonment. Operations Research **59**(4), 835–846 (2011)
22. Liu, Y., Whitt, W.: Large-time Asymptotics for the Gt/Mt/st+GIt Many-server Fluid Queue with Abandonment. Queueing Syst. Theory Appl. **67**(2), 145–182 (2011). DOI 10.1007/s11134-010-9208-8
23. Liu, Y., Whitt, W.: Nearly periodic behavior in the overloaded $G/D/s + GI$ queue. Stoch. Syst. **1**(2), 340–410 (2011)
24. Liu, Y., Whitt, W.: A many-server fluid limit for the queueing model experiencing periods of overloading. Operations Research Letters **40**(5), 307 – 312 (2012)
25. Liu, Y., Whitt, W.: Algorithms for Time-Varying Networks of Many-Server Fluid Queues. INFORMS Journal on Computing **26**(1), 59–73 (2014)
26. Liu, Y., Whitt, W.: Many-server heavy-traffic limit for queues with time-varying parameters. The Annals of Applied Probability **24**(1), 378–421 (2014)
27. Long, Z., Zhang, J.: Convergence to equilibrium states for fluid models of many-server queues with abandonment. Operations Research Letters **42**(6), 388–393 (2014)
28. Mandelbaum, A., Massey, W.A., Reiman, M.I.: Strong approximations for Markovian service networks. Queueing Systems **30**(1-2), 149–201 (1998)
29. Mandelbaum, A., Massey, W.A., Reiman, M.I., Stolyar, A.L., Rider, B.: Queue Lengths and Waiting Times for Multiserver Queues with Abandonment and Retrials. Telecommunication Systems **21**(2), 149–171 (2002)
30. Mandelbaum, A., Momčilović, P.: Queues with many servers and impatient customers. Mathematics of Operations Research **37**(1), 41–65 (2012). DOI 10.1287/moor.1110.0530. URL http://dx.doi.org/10.1287/moor.1110.0530
31. Meyn, S.P., Tweedie, R.L.: Stability of Markovian Processes III: Foster-Lyapunov Criteria for Continuous-Time Processes. Advances in Applied Probability **25**(3) (1993)
32. Nelson, Barry L. and Taaffe, Michael R.: The [Pht/Pht/8]K Queueing System: Part II–The Multiclass Network. INFORMS J. on Computing **16**(3), 275–283 (2004). DOI 10.1287/ijoc.1040.0071
33. Pavlov, A., Wouw, N.V.D., Nijmeijer, H.: Convergent piecewise affine systems: analysis and design Part I: continuous case. In: 44th IEEE Conference on Decision and Control and European Control Conference ECC 2005 (2005)
34. S. Boyd and L. El Ghaoui and E. Feron and V. Balakrishnan: Linear Matrix Inequalities in System and Control Theory, *Studies in Applied Mathematics*, vol. 15. SIAM, Philadelphia, PA (1994)
35. Stewart, W.J.: Probability, Markov Chains, Queues, and Simulation. Princeton University Press (2009)
36. Urgaonkar, B., Pacifici, G., Shenoy, P., Spreitzer, M., Tantawi, A.: An analytical model for multi-tier internet services and its applications. In: SIGMETRICS, pp. 291–302 (2005)
37. Van Houdt, B.: A mean field model for a class of garbage collection algorithms in flash-based solid state drives. In: SIGMETRICS, pp. 191–202. ACM, New York, NY, USA (2013)
38. Van Houdt, B., Bortolussi, L.: Fluid limit of an asynchronous optical packet switch with shared per link full range wavelength conversion. In: SIGMETRICS, pp. 113–124 (2012)
39. Whitt, W.: Fluid Models for Multiserver Queues with Abandonments. Operations Research **54**(1), 37–54 (2006)