Insensitivity to Service-time Distributions for Fluid Queueing Models

Max Tschaikowski
Electronics and Computer Science
University of Southampton, UK
m.tschaikowski@soton.ac.uk

Mirco Tribastone
Electronics and Computer Science
University of Southampton, UK
m.tribastone@soton.ac.uk

ABSTRACT

We study fluid limits based on ordinary differential equations (ODEs) for Markovian queueing models where nonexponential service times are fit by appropriate Coxian distributions to match their first and second moments. We focus on a heavy-load regime, whereby the fluid limit of the queue-length process of the nonexponential queue estimates a bottleneck situation. Under this condition, we show that the ODE solution admits a steady state which is insensitive to the service-time distribution: The ODE steady state only depends on the mean service times. By contrast, the steady-state average performance measures computed by Markovian analysis are in general dependent on the higher-order moments of the service-time distribution. A numerical investigation shows that, given any two Markovian queueing models with Coxian-distributed service times with the same mean and different variance, the model with lower variance converges more rapidly to the (same) fluid limit than the one with higher variance.

Keywords

Queueing models, fluid limits, Coxian distributions, insensitivity

1. INTRODUCTION

Over the past few years fluid models have gained increased popularity as an effective tool for the performance evaluation of large-scale systems. They have been used, for instance, to study a wide range of distributed and networked systems, including load balancing [12, 6], optical switches [14], virtualized environments [1], and peer-to-peer networks [18].

The theory is well established for Markov population processes, i.e., continuous-time Markov chains (CTMCs) where the state descriptor is a vector of nonnegative integers that tracks the populations of different kinds of agents in the system under scrutiny [8]. Under relatively mild assumptions, a Markov population process admits a *fluid limit* as the solution to a system of ordinary differential equations (ODEs), which is interpreted as the asymptotic sample path of the Markov process when the populations are infinite [8]. From a practical point of view, this can be seen as a deterministic approximation that becomes increasingly accurate as the populations get larger. There is a clear advantage in using this approximation: The computational cost of the ODE analysis (typically performed using numerical integration) is roughly independent of the population sizes in the system, unlike CTMC analysis (either conducted by numerical solution via linear algebra or by simulation, see e.g., [13]), which instead is well-known to suffer to a great extent from the curse of dimensionality.

Being based on a Markov process, a crucial — and perhaps limiting — modeling assumption in developing fluid models is that the activities are distributed exponentially. However, there is substantial evidence that, in many cases of practical relevance, the distributions of certain realistic durations are not exponential. In some cases there are activities that may be well represented by deterministic durations (for instance, in the case of timers), while in other cases the distributions feature heavy tails, e.g., [5].

It is well-known that nonexponential durations can be expressed as appropriate expansions into a number of exponential stages in such a way that the resulting stochastic process is still Markov, thus amenable to the same CTMC analysis. In this paper we consider two specific classes expansions which, depending on the squared coefficient of variation (hereafter denoted by SCV or $\mathcal{V}$) of a given nonexponential activity of interest, can be used for fitting its first and second moments by a Coxian distribution [2].

Our main purpose is to study the impact on fluid analysis of models which have such Coxian distributions. We carry out this investigation in a queueing-theoretic context, by analyzing a tandem network where a multi-server queue with Coxian-distributed service times is fed by an exponential *delay station*, i.e., a queue with infinite servers (see Figure 1). We provide a result of insensitivity of the fluid steady state, in essence by showing that any two Coxian distributions with the same means but different SCVs will have *the same* fluid steady-state queue lengths. Interestingly, this result holds even between two Coxian distributions with different number of stages (which lead to ODE systems of different size). As a byproduct, the steady-state queue lengths are also equal to those of a tandem network when the Coxian distribution degenerates to an exponential one with the same mean. In all cases, insensitivity occurs *only* in the steady state, while the fluid transient behaviors of the queue lengths are not comparable in general.

ValueTools'13, December 10 – 12 2013, Turin, Italy

VALUETOOLS 2013, December 10-12, Torino, Italy

DOI 10.4108/icst.valuetools.2013.254374

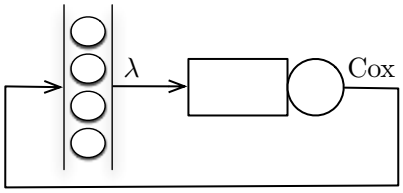

Figure 1: Tandem queueing network consisting of an infinite-server exponential queue (with rate λ) followed by a multi-server queue with Coxian-distributed service times.

We are able to obtain this result by studying the system in a *heavy-load regime* (HLR). Roughly speaking, it corresponds to assuming that in the *fluid* solution the Coxian queue is the bottleneck, i.e., there are more jobs than available servers. Insensitivity is proven to hold whenever this regime is observed in any two networks featuring distinct Coxian distributions with the same mean. Our HLR condition is somewhat analogous to the *heavy-traffic* regime that has been extensively studied in the literature, where a queueing system is considered in the limit behavior of its utilization *approaching 1* (e.g., [16]). It is, however, different since the fluid solution is attained only if the populations of customers *and* servers tend to infinity.

We find that if the network is in HLR then the fluid solution estimates a steady-state utilization for the bottleneck queue of *exactly 1*, i.e., all servers are busy with probability 1, even if in each queueing network of the limiting sequence the population of servers is allowed to scale up directly proportionally to the population of clients. This accounts for the different behavior between the fluid model, which is insensitive, and the stochastic model, which is not, because the latter will tend to unitary utilization in the limit when the client populations go to infinity.

Overall, this investigation tells us that two queueing networks with different service-time distributions may not be told apart by their steady-state fluid analyses, if these systems are characterized by the same means. Especially when interpreting the fluid limit as an approximate estimate of a large-scale — but finite — system, it then becomes a natural question to ask which kinds of distributions are generally approximated more effectively in practice. To this end, we carry out an numerical investigation on the CTMC analysis. This analysis shows that, fixing a queueing network with the same average service times and the same populations of clients and servers, the network with smaller SCVs for the service times is consistently better approximated from the fluid solution that the network with higher SCVs.

Related Work. In stochastic queueing networks, insensitivity has been an intense area of research ever since the pioneering work of Erlang. In his loss model (a system with Poisson arrivals, generally distributed parallel servers, and no waiting room), the performance measures are insensitive to the service-time distribution (e.g., [4]). Whittle relates insensitivity to partial-balance equations of Markov processes [17]. Schassberger [11] and Miyazawa [10] relate insensitivity to decomposability in product form for closed queueing networks. These approaches cannot be applied in our context because stochastic insensitivity does not hold here. In fact, we will show that two models with the same mean service times but with different variances will in general provide different stochastic steady-state measures, but the same fluid estimates. Instead, the result of Bramson et. al [3] is similar in spirit to ours, but is carried out at the stochastic level for an open tandem network with Poisson arrival rates. To sum up, none of these approaches is applicable to show our result of insensitivity.

Paper Organization. Section 2 covers the mathematical background, introducing fitting methods by Coxian distributions, the stochastic model of the queueing network in Fig. 1 and its fluid approximation. It concludes with an illustrating numerical example that shows insensitivity of the fluid steady state but dependence of the stochastic steady states from higher order moments than the mean of the service-time distributions. Section 3 proves the general results of insensitivity of the fluid steady states under the HLR assumption, and relates this to the utilization of the bottleneck queue. Section 4 presents the numerical evaluation to study the speed of convergence to the fluid steady states for queueing networks with different Coxian distributions with the same mean. Section 5 concludes with some final remarks. Proofs of Propositions are reported in the appendix.

2. PRELIMINARIES

To make the paper self-contained, here we overview the preliminary material required to set the context.

In this paper we use the following notation. Greek letters λ and μ are positive real numbers that denote rates of exponential distributions. In particular, λ is used for the rate of the delay station while μ is used for Coxian distributions. The SCV of a random variable (here, always a service-time distribution) is the ratio between its variance and its squared mean. Vectors are denoted by bold faced symbols. ODEs will always be autonomous, and are given in the form $\dot{\boldsymbol{x}} = f(\boldsymbol{x})$, where $\boldsymbol{x}$ is the vector of dependent variables, functions of time t, $\dot{\boldsymbol{x}}$ is the derivative of $\boldsymbol{x}$ with respect to t, and f is the given vector field. We denote a Coxian distribution by $\mathrm{Cox}(\boldsymbol{\mu}, \boldsymbol{r})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ is a vector with $1/\mu_i$ being the average holding time in each stage, and $\boldsymbol{r} = (r_1, r_2, \ldots, r_{n-1})$, where r_i, $1 \leq i \leq n-1$, indicates the probability that a job in the i-th stage of service goes to the $(i+1)$-th stage. Thus, an Erlang-3 distribution with average service time $1/\mu$ is denoted by $\mathrm{Cox}((3\mu, 3\mu, 3\mu), (1, 1))$.

More in general, Coxian distributions can be used to fit the first two moments of any random variable with given finite mean $1/\mu_0$ and $\mathcal{V}$. We use two distinct well-known fittings (see, e.g., [2]) according to whether $\mathcal{V} \geq 1$ (high SCV) or $\mathcal{V} < 1$ (low SCV).

High SCV ($\mathcal{V} \geq 1$). In this case, the Coxian distribution has two stages, defined as $\mathrm{Cox}((\mu_h, p_h\mu_h), p_h)$ with

$$\mu_h = 2\mu_0, \qquad p_h = \frac{1}{2\mathcal{V}}.$$

We let $q_h := 1 - p_h$.

Low SCV ($\mathcal{V} < 1$). In this case, the Coxian distribution is given by $\mathrm{Cox}((\mu_l, \ldots, \mu_l), (p_l, 1, 1, \ldots, 1))$, with

$$\begin{aligned} n &= \left\lceil \frac{1}{\mathcal{V}} \right\rceil, \\ p_l &= 1 - \frac{2n \cdot \mathcal{V} + (n-2) - \sqrt{n^2 + 4 - 4n \cdot \mathcal{V}}}{2(\mathcal{V}+1)(n-1)}, \\ \mu_l &= \mu_0\big((1-p_l) + np_l\big). \end{aligned} \tag{1}$$

Notice that any Erlang-n distribution is also covered, by setting $p_l = 1$.

2.1 Queueing Network Model

As discussed in Section 1, we consider a tandem queueing network with infinite buffers where a multi-server queue with Coxian service times is fed by an infinite-server queue (the *delay station*) with exponential service times. Let the Coxian distribution be denoted by

$$\mathrm{Cox}((\mu_1, \ldots, \mu_n), (r_1, r_2, \ldots, r_{n-1})),$$

and let λ be the exponential rate. The network is a CTMC with state descriptor given as the vector

$$\boldsymbol{\omega} = (c_0, c_1, \ldots, c_n, s_1, \ldots, s_n),$$

where each element is a nonnegative integer, with the following meaning: c_0 gives the population of jobs in the delay station; c_i gives the population of jobs in the i-th stage of the Coxian, while s_i gives the population of servers in the i-th stage of the Coxian, with $1 \leq i \leq n$. Therefore the queue-length process at the delay station is given directly by c_0 while the queue-length process at the Coxian queue is given by the sum $c_1 + \ldots + c_n$; this metric counts the number of jobs waiting as well as those which are in service.

Let $\mathbb{1}_p$ be a zero-valued vector of length $2n+1$, except for its element at position p which is equal to 1, with $p \in \{c_0, \ldots, c_n, s_1, \ldots, s_n\}$. The transition rates of the CTMC are given in terms of a set of real-valued functions of the state descriptor, each associated with a *jump vector*. Each function gives the rate to the state computed by adding the jump vector to the current state. In this model, these functions are denoted by $f_{0,1}$, $f_{j,j+1}$ and $f_{j,0}$ for all $1 \leq j \leq n-1$. Their corresponding jump vectors are denoted by $\Delta\boldsymbol{\omega}_{0,1}$, $\Delta\boldsymbol{\omega}_{j,j+1}$, and $\Delta\boldsymbol{\omega}_{j,0}$, respectively. Specifically,

$$f_{0,1}(\boldsymbol{\omega}) = \lambda c_0, \qquad \Delta\boldsymbol{\omega}_{0,1} = -\mathbb{1}_{c_0} + \mathbb{1}_{c_1}$$

models the exponential rate at the delay station. Upon service, one job goes into the first stage of the Coxian service. The function

$$f_{n,0}(\boldsymbol{\omega}) = \mu_n \min(c_n, s_n), \quad \Delta\boldsymbol{\omega}_{n,0} = \mathbb{1}_{c_0} - \mathbb{1}_{c_n} + \mathbb{1}_{s_1} - \mathbb{1}_{s_n}$$

models the end of service at the n-th stage of the Coxian. Finally, the other functions are as follows.

$$\begin{aligned} f_{j,j+1}(\boldsymbol{\omega}) &= r_j \mu_j \min(c_j, s_j), \\ &\qquad \Delta\boldsymbol{\omega}_{j,j+1} = \mathbb{1}_{c_{j+1}} - \mathbb{1}_{c_j} + \mathbb{1}_{s_{j+1}} - \mathbb{1}_{s_j}, \\ f_{j,0}(\boldsymbol{\omega}) &= (1-r_j)\mu_j \min(c_j, s_j), \\ &\qquad \Delta\boldsymbol{\omega}_{j,0} = \mathbb{1}_{c_0} - \mathbb{1}_{c_j} + \mathbb{1}_{s_1} - \mathbb{1}_{s_j}, \end{aligned}$$

for all $1 \leq j \leq n-1$. The functions $f_{j,j+1}$ describe a job in stage j which goes into the next stage $j+1$. Accordingly, this will decrement the populations of jobs and servers in stage j and increment the respective populations in stage $j+1$. The rate $\mu_j \min(c_j, s_j)$ is the overall rate at which jobs are served in stage j; this comes from the assumption of i.i.d. exponential service times with rate μ_j. The functions $f_{j,0}$ describe a job in stage j finishes service and goes into the delay station. Hence, the increment of the population of c_0 and of s_1, the latter because one of the servers will be ready to serve a new customer from stage 1.

Example 1 (High-SCV Model). *Consider a tandem queueing network where the Coxian-distributed queue has high SCV. Then, we have $n = 2$ and $\boldsymbol{\omega} = (c_0, c_1, c_2, s_1, s_2)$. The functions are*

$$\begin{aligned} f_{0,1}(\boldsymbol{\omega}) &= \lambda c_0, & \Delta\boldsymbol{\omega}_{0,1} &= \mathbb{1}_{c_1} - \mathbb{1}_{c_0}, \\ f_{1,2}(\boldsymbol{\omega}) &= p_h \mu_h \min(c_1, s_1), & \Delta\boldsymbol{\omega}_{1,2} &= \mathbb{1}_{c_2} - \mathbb{1}_{c_1} + \mathbb{1}_{s_2} - \mathbb{1}_{s_1}, \\ f_{1,0}(\boldsymbol{\omega}) &= q_h \mu_h \min(c_1, s_1), & \Delta\boldsymbol{\omega}_{1,0} &= \mathbb{1}_{c_0} - \mathbb{1}_{c_1} \\ f_{2,0}(\boldsymbol{\omega}) &= p_h \mu_h \min(c_2, s_2), & \Delta\boldsymbol{\omega}_{2,0} &= \mathbb{1}_{c_0} - \mathbb{1}_{c_2} + \mathbb{1}_{s_1} - \mathbb{1}_{s_2}. \end{aligned}$$

The transition functions for a tandem queueing network with a Coxian-distributed queue with low-SCV model can be obtained similarly.

We use this CTMC representation to introduce the fluid limit result, as discussed next.

2.2 Fluid Limits

With the transition functions, a concrete CTMC is completely characterized by fixing an initial state, hereafter denoted by $\boldsymbol{\omega}(0) \in \mathbb{N}_{\geq 0}^{2n+1}$. In the previous example, setting $\boldsymbol{\omega}(0) = (0, C, 0, S, 0)$ with $C, S > 0$, means that there are C jobs in the system, all waiting for service at the Coxian queue. Moreover, the total number of servers in the system is S and, similarly to the jobs, all of them are in the first stage of the Coxian. From this, the whole CTMC state space may be systematically derived by applying all nonzero valued functions to the current state, until no further states can be found. For instance, from $\boldsymbol{\omega}(0)$, the chain can go into the states $(0, C-1, 1, S-1, 1)$ and $(1, C-1, 0, S, 0)$ with rates $p_h \mu_h \min(C, S)$ and $q_h \mu_h \min(C, S)$, respectively.

Given an initial vector $\boldsymbol{\omega}(0)$, it is also possible to construct a family of CTMCs indexed by a single parameter, denoted $v \in \mathbb{N}$. The v-th element of this sequence has initial state denoted by $\boldsymbol{\omega}_v(0)$, such that

$$\boldsymbol{\omega}_v(0) = v \cdot \boldsymbol{\omega}(0), \qquad \text{for all } v.$$

In the example above, we have that $\boldsymbol{\omega}_1(0) = (0, C, 0, S, 0)$, $\boldsymbol{\omega}_2(0) = (0, 2C, 0, 2S, 0)$, and so forth. Let $\{\boldsymbol{X}_v(t), v \in \mathbb{N}\}$ denote such a family of CTMCs.

Let us now consider the *density process* $\boldsymbol{X}_v(t)/v$. Such a process will make transitions of order $O(1/v)$ with rates of order $O(v)$; that is, the density process will make increasingly smaller jumps at increasingly faster rates with v. This suggests a continuous limit behavior as $v \to \infty$. Indeed, using a celebrated result from Kurtz [8], it can be shown that, for any finite t, $\boldsymbol{X}_v(t)/v$ will converge in probability to

$$\boldsymbol{x}(t) := \big(C_0(t), \ldots, C_n(t), S_1(t), \ldots, S_n(t)\big), \tag{2}$$

the unique solution to the (nonlinear) ODE system

$$\begin{aligned} \dot{\boldsymbol{x}} = \Delta\boldsymbol{\omega}_{0,1} f_{0,1}(\boldsymbol{x}) &+ \\ &+ \sum_{j=1}^{n} \Delta\boldsymbol{\omega}_{j,0} f_{j,0}(\boldsymbol{x}) + \sum_{j=1}^{n-1} \Delta\boldsymbol{\omega}_{j,j+1} f_{j,j+1}(\boldsymbol{x}), \end{aligned} \tag{3}$$

subject to initial condition $\boldsymbol{x}(0) = \boldsymbol{\omega}(0)$.

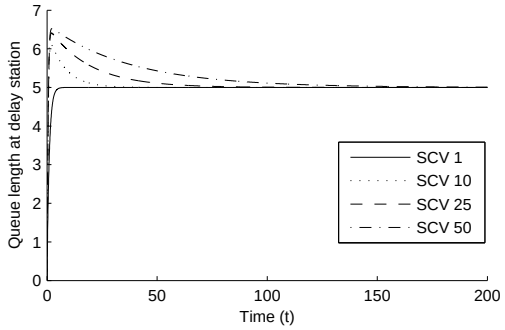

Figure 2: Insensitivity of the fluid steady-state queue length of Example 1 for Coxian distributions with the same mean, $1/\mu_0 = 1.0$, and different SCVs.

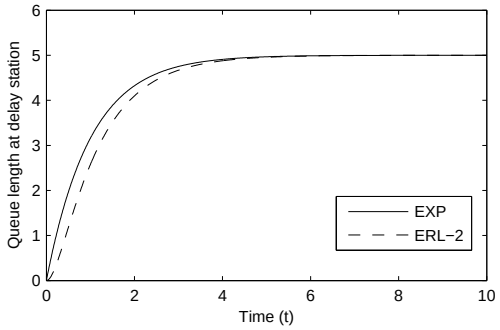

Figure 3: Insensitivity of the fluid steady-state queue length for service times with exponential and Erlang-2 distributions.

Pragmatically, this allows us to introduce the approximation $\boldsymbol{X}_v(t) \approx v\boldsymbol{x}(t)$ for large v (hence, for large populations).

In Example 1, let us write $\boldsymbol{x} = (C_0, C_1, C_2, S_1, S_2)$. Then, the ODE system is (component-wise) given by

$$\begin{aligned}
\dot{C}_0 &= q_h\mu_h \min(C_1, S_1) + p_h\mu_h \min(C_2, S_2) - \lambda C_0 \\
\dot{C}_1 &= \lambda C_0 - \mu_h \min(C_1, S_1) \\
\dot{C}_2 &= p_h\mu_h \min(C_1, S_1) - p_h\mu_h \min(C_2, S_2) \\
\dot{S}_1 &= p_h\mu_h \min(C_2, S_2) - p_h\mu_h \min(C_1, S_1) \\
\dot{S}_2 &= p_h\mu_h \min(C_1, S_1) - p_h\mu_h \min(C_2, S_2)
\end{aligned} \tag{4}$$

2.3 Illustrating Example

We are now ready to provide an illustrating example of steady-state insensitivity of the fluid analysis. Let us consider Example 1 and its ODE system (4) with four different parameterizations, in order to obtain Coxian-distributed service times with the same mean, $1/\mu_0 = 1.0$, but different SCVs, set to 1, 10, 25, and 50, respectively. Figure 2 shows the fluid trajectories for $C_0(t)$ when all models have $\lambda = 1.0$ and initial condition $\boldsymbol{x}(0) = (0, 10, 0, 5, 0)$. It can be seen that the transient behavior does depend on the SCV, however all trajectories tend to a steady state of 5.000. Since the network is closed, i.e., $C_0(t) + C_1(t) + C_2(t) = 10$ for all t, then also the steady-state queue length of the Coxian queue (computed, for all t as the sum $C_1(t) + C_2(t)$) is insensitive to the SCV. Thus, we have shown an example of insensitivity across Coxian distributions belonging to the same class with $\mathcal{V} \geq 1$.

Let us now consider two further models. The first one a is Coxian distribution with $\mathcal{V} < 1$. In particular, let us study the Erlang-2 case, whose ODE system is given by (4) by setting $p_h = 1.0$ and $\mu_h = 2\mu_0 = 2.0$. The second model is instead a Coxian queue that degenerates to an exponential one with rate μ_0. Its ODE system is given by

$$\begin{aligned}
\dot{C}_0 &= \mu_0 \min(C_1, S_1) - \lambda C_0 \\
\dot{C}_1 &= \lambda C_0 - \mu_0 \min(C_1, S_1) \\
\dot{S}_1 &= 0
\end{aligned} \tag{5}$$

where the last equation indicates that the server never changes state, hence $S_1(t)$ is equal to its initial condition for all t.

Figure 3 shows the fluid trajectories for these two models with initial conditions $(0, 10, 0, 5, 0)$ and $(0, 10, 5)$, hence maintaining the same total populations of jobs and servers in both networks. Again, it can be noticed that while the transient behaviors are in general different, the steady states coincide, and are also equal to those attained by the high-SCV Coxian distributions in the previous example. Thus, at least in this case, we showed that the fluid steady state is insensitive to the SCV of the service-time distribution. However, in general the networks are *stochastically different*, in that the steady-state average queue lengths computed by CTMC analysis *are not* insensitive, as shown in Table 1. (Although SCV 1 and EXP are an exception, notice that they have the same first two moments.) Interestingly, the CTMC results are ordered by increasing SCV of the service-time distribution. In this case, we observe that the higher the SCV the further away the CTMC steady-state is from the fluid estimate (equal to 5.000). This phenomenon will be numerically analyzed in more detail in Section 4.

ERL-2	EXP	SCV 1	SCV 10	SCV 25	SCV 50
4.537	4.504	4.504	4.429	4.420	4.416

Table 1: CTMC steady-state analysis. Top line: model name, cf. legends of Figures 2 and 3; bottom line: average queue length of the delay station. The steady-state fluid estimate is equal to 5.000.

In the next section we show that, indeed, the fluid state states are insensitive to higher-order moments than the mean service times in our tandem queueing network.

3. ANALYSIS OF FLUID STEADY STATES

The analysis of the steady states of the ODE system (3) is made difficult by the fact that the system is nonlinear; indeed, it is piece-wise linear. In this case, the nonlinearities are due to the presence of a minimum function for each stage of the Coxian. Thus the solution space $\mathbb{R}^{2n+1}$ can be partitioned in 2^n polytopes, where each polytope contains some linear system. For instance, the ODE system (4) for the 2-stage high-SCV Coxian distribution can be represented in terms of $2^2 = 4$ linear systems, as shown in Figure 4. It is evident that (3) may admit up to 2^n fluid steady states, one for each linear ODE system.

However, if we are able to show that for certain initial conditions the corresponding solution of the piece-wise linear system (3) remains within the same polytope and converges

$S_1 > C_1, \quad S_2 > C_2:$	$S_1 > C_1, \quad S_2 \le C_2:$	$S_1 \le C_1, \quad S_2 > C_2:$	$S_1 \le C_1, \quad S_2 \le C_2:$
$\dot{C}_0 = q_h\mu_h C_1 + p_h\mu_h C_2 - \lambda C_0$	$\dot{C}_0 = q_h\mu_h C_1 + p_h\mu_h S_2 - \lambda C_0$	$\dot{C}_0 = q_h\mu_h S_1 + p_h\mu_h C_2 - \lambda C_0$	$\dot{C}_0 = q_h\mu_h S_1 + p_h\mu_h S_2 - \lambda C_0$
$\dot{C}_1 = \lambda C_0 - \mu_h C_1$	$\dot{C}_1 = \lambda C_0 - \mu_h C_1$	$\dot{C}_1 = \lambda C_0 - \mu_h S_1$	$\dot{C}_1 = \lambda C_0 - \mu_h S_1$
$\dot{C}_2 = p_h\mu_h C_1 - p_h\mu_h C_2$	$\dot{C}_2 = p_h\mu_h C_1 - p_h\mu_h S_2$	$\dot{C}_2 = p_h\mu_h S_1 - p_h\mu_h C_2$	$\dot{C}_2 = p_h\mu_h S_1 - p_h\mu_h S_2$
$\dot{S}_1 = p_h\mu_h C_2 - p_h\mu_h C_1$	$\dot{S}_1 = p_h\mu_h S_2 - p_h\mu_h C_1$	$\dot{S}_1 = p_h\mu_h C_2 - p_h\mu_h S_1$	$\dot{S}_1 = p_h\mu_h S_2 - p_h\mu_h S_1$
$\dot{S}_2 = p_h\mu_h C_1 - p_h\mu_h C_2$	$\dot{S}_2 = p_h\mu_h C_1 - p_h\mu_h S_2$	$\dot{S}_2 = p_h\mu_h S_1 - p_h\mu_h C_2$	$\dot{S}_2 = p_h\mu_h S_1 - p_h\mu_h S_2$

Figure 4: The four linear ODE systems representing the piece-wise linear ODE (4).

to an equilibrium point, then the fluid steady state can be inferred by solving a linear system of equations. In order to achieve this, we make the assumption of a *heavy-load regime* (HLR). Let C and S be the total number of jobs and servers in the network. Using the state representation (2), we have

$$C = \sum_{i=0}^{n} C_i \qquad \text{and} \qquad S = \sum_{i=1}^{n} S_i.$$

Our HLR assumption requires that C is larger than S. More specifically, for any given rate parameterization of the network and any given S, we find that there always exists a sufficiently large $\tilde{C}$ such that, for all $C \ge \tilde{C}$, the network behaves in such a way that $C_i(t) \ge S_i(t)$ holds for all $t \ge 0$ and $1 \le i \le n$ (cf. Proposition 3). For instance, in (4) the HLR assumption reduces to analyzing the case $C_1 \ge S_1, C_2 \ge S_2$ with an initial condition $(0, C, S, 0)$, for C large enough. Analytically, the HLR allows us to study the nonlinear system (3) using techniques of linear ODE systems, since (3) starts in and never leaves one of its regions where it is indeed linear.

In the remainder, we consider the cases for the Coxian distribution with low SCV and high SCV separately.

3.1 Low-SCV Coxian Distribution

Let us consider the tandem queueing network in Figure 1. Let the service rate for the exponential delay station be given by λ, and let the Coxian-distributed queue be characterized by the n-stage low-SCV Coxian fit of a given distribution with mean $1/\mu_0$ and $\mathcal{V} < 1$. Then, according to Sections 2.1 and 2.2, the following (linear) ODE system represents the fluid limit of the network in the case of $C_i \ge S_i$ for all $1 \le i \le n$.

$$\begin{aligned}
\dot{C}_0 &= (1-p_l)\mu_l S_1 + \mu_l S_n - \lambda C_0 \\
\dot{C}_1 &= \lambda C_0 - \mu_l S_1 & \dot{S}_1 &= \mu_l S_n - \mu_l p_l S_1 \\
\dot{C}_2 &= \mu_l p_l S_1 - \mu_l S_2 & \dot{S}_2 &= \mu_l p_l S_1 - \mu_l S_2 \\
\dot{C}_3 &= \mu_l S_2 - \mu_l S_3 & \dot{S}_3 &= \mu_l S_2 - \mu_l S_3 \\
&\vdots & &\vdots \\
\dot{C}_n &= \mu_l S_{n-1} - \mu_l S_n & \dot{S}_n &= \mu_l S_{n-1} - \mu_l S_n
\end{aligned} \tag{6}$$

Let this system be represented in matrix notation as

$$\dot{\boldsymbol{x}} = M_l^n \boldsymbol{x}, \qquad \text{with} \quad M_l^n = \left(\begin{array}{c|c} A_n & B_n \\ \hline 0 & D_n \end{array}\right)$$

where

$$\boldsymbol{x} = (C_0, C_1, \ldots, C_n, S_1, \ldots, S_n)$$

and $M_l^n \in \mathbb{R}^{(2n+1)\times(2n+1)}$ is given by

$$\left(\begin{array}{cccc|cccccc}
-\lambda & 0 & \ldots & 0 & (1-p_l)\mu_l & 0 & 0 & 0 & \ldots & 0 & \mu_l \\
\lambda & 0 & \ldots & 0 & -\mu_l & 0 & 0 & 0 & \ldots & 0 & 0 \\
0 & 0 & \ldots & 0 & \mu_l p_l & -\mu_l & 0 & 0 & \ldots & 0 & 0 \\
0 & 0 & \ldots & 0 & 0 & \mu_l & -\mu_l & 0 & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & \mu_l & -\mu_l \\
\hline
0 & 0 & \ldots & 0 & -\mu_l p_l & 0 & 0 & 0 & \ldots & 0 & \mu_l \\
0 & 0 & \ldots & 0 & \mu_l p_l & -\mu_l & 0 & 0 & \ldots & 0 & 0 \\
0 & 0 & \ldots & 0 & 0 & \mu_l & -\mu_l & 0 & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & \mu_l & -\mu_l
\end{array}\right).$$

It is well-known that the solution of a linear ODE system is closely related to the eigenvalues of the underlying matrix. First, we study the nature of the eigenvalues of M_l^n.

PROPOSITION 1. *The complex number $0 \in \mathbb{C}$ is an eigenvalue of $M_l^n \in \mathbb{R}^{(2n+1)\times(2n+1)}$ with geometric and algebraic multiplicity equal to $n+1$. Apart from 0, M_l^n has only eigenvalues with negative real parts.*

The above result is preliminary toward proving the existence of a fluid steady state for (6). Additionally, we can estimate the corresponding speed of convergence.

PROPOSITION 2. *Let us assume that λ, μ_l, p_l are fixed. Then, there exist $a, b_1, b_2 > 0$ such that, for all $C, S > 0$, there is an $\boldsymbol{x}_{ss} \in \mathbb{R}^{2n+1}$ such that the solution of (6) subject to $\boldsymbol{x}(0) = C\mathbb{1}_{C_1} + S\mathbb{1}_{S_1}$ can be written as $\boldsymbol{x}(t) = \boldsymbol{x}_{ss} + \boldsymbol{x}_{tr}(t)$ with*

$$\|\mathbf{x}_{tr}(t)\|_\infty \le (b_1 C + b_2 S) e^{-at}(1+t)^n, \quad \textit{for all} \quad t \ge 0.$$

As shown below, the last two propositions imply that the inequalities $C_i(t) \ge S_i(t)$, with $1 \le i \le n$, remain valid for all $t \ge 0$. As a consequence, the piecewise linear system (3) coincides with the linear system (6).

PROPOSITION 3. *Let $\boldsymbol{x}$ denote the solution of (6) with $\boldsymbol{x}(0) = C\mathbb{1}_{C_1} + S\mathbb{1}_{S_1}$ for arbitrary but fixed λ, μ_l, p_l, S. Then there exists a $\tilde{C} > 0$ such that for all $C \ge \tilde{C}$ it holds that*

$$C_i(t) \ge S_i(t), \quad \textit{for all} \quad 1 \le i \le n \quad \textit{and} \quad t \ge 0.$$

Crucially, $\boldsymbol{x}$ is also the solution of (3) with low-SCV Coxian service time given by (1), for the initial condition $\boldsymbol{x}(0)$.

Using the former propositions, we are in a position to provide the steady state of (3) under HLR.

THEOREM 1. *For any given $\lambda, S > 0$, there exists a $\tilde{C} > 0$ such that for all $C \geq \tilde{C}$ the fluid approximation (3) with low-SCV Coxian service time given by (1), subjected to $\boldsymbol{x}(0) = C\mathbb{1}_{C_1} + S\mathbb{1}_{S_1}$, has the following steady state:*

$$C_0 = \frac{\mu_0}{\lambda} S$$

$$C_1 = C - \sum_{0 \leq i \leq n, i \neq 1} C_i = C - S\left(\frac{\mu_0}{\lambda} + \frac{(n-1)p_l}{1+(n-1)p_l}\right),$$

$$C_i = \frac{p_l S}{1+(n-1)p_l}, \quad i = 2, \ldots, n,$$

$$S_1 = \frac{S}{1+(n-1)p_l},$$

$$S_i = C_i, \quad i = 2, \ldots, n.$$

PROOF. Thanks to Proposition 3, all one has to do is to solve $M_l^n \boldsymbol{x} = \mathbf{0}$ in the case of $\sum_{i=1}^n S_i = S$ and $\sum_{i=0}^n C_i = C$. This gives the stated values of $S_1, \ldots, S_n$ and $C_2, \ldots, C_n$. For C_0, instead, one gets $C_0 = \frac{\mu_l}{\lambda} \frac{S}{1+(n-1)p_l}$ which yields the claim since $\mu_l = \mu_0\big((1-p_l) + np_l\big)$, as discussed in Section 2. □

From this result we immediately have that, under the HLR assumption, any two queueing networks with the same mean service times will have identical steady-state queue lengths. This is because the steady-state queue length of the delay station, C_0, only depends on the means, and the steady-state queue length of the Coxian-distributed queue is given by $C - C_0$.

3.2 High-SCV Coxian Distribution

In this section we develop the analogous result for the tandem queueing network in Figure 1, where delay station is still exponential with rate λ and the Coxian-distributed queue has mean service time $1/\mu_0$. However, here we study the two-stage high-SCV Coxian fit, whose ODE system has been shown to be (4). Again, let us impose our HLR assumption, i.e., $C_i \geq S_i$, for $i = 1, 2$. Then, the resulting linear ODE system can be written in the matrix notation

$$\dot{\boldsymbol{x}} = M_h \boldsymbol{x}$$

where $\boldsymbol{x} = (C_0, C_1, C_2, S_1, S_2)$ and $M_h \in \mathbb{R}^{5\times 5}$ is given by

$$M_h = \begin{pmatrix} -\lambda & 0 & 0 & q_h\mu_h & p_h\mu_h \\ \lambda & 0 & 0 & -\mu_h & 0 \\ 0 & 0 & 0 & p_h\mu_h & -p_h\mu_h \\ 0 & 0 & 0 & -p_h\mu_h & p_h\mu_h \\ 0 & 0 & 0 & p_h\mu_h & -p_h\mu_h \end{pmatrix}.$$

By performing a similar argumentation to the case of low-SCV Coxian distributions, we have the following result.

THEOREM 2. *For any given $\lambda, S > 0$, there exists a $\tilde{C} > 0$ such that for all $C \geq \tilde{C}$ the fluid approximation (4) subjected to $\boldsymbol{x}(0) = C\mathbb{1}_{C_1} + S\mathbb{1}_{S_1}$ has the following steady state:*

$$C_0 = \frac{\mu_0}{\lambda} S,$$

$$C_1 = C - (C_0 + C_2) = C - S\left(\frac{\mu_0}{\lambda} + \frac{1}{2}\right),$$

$$C_2 = \frac{S}{2},$$

$$S_1 = S_2 = \frac{S}{2}.$$

PROOF. It is clear that one can apply the very same proof strategy as in Proposition 2 and 3, if one is able to show that, apart from 0, all eigenvalues of M_h have negative real parts and that the algebraic and geometric multiplicity of 0 are the same. To this see, we note that

$$\wp_{M_h}(z) = \begin{vmatrix} z+\lambda & 0 & 0 \\ -\lambda & z & 0 \\ 0 & 0 & z \end{vmatrix} \cdot \begin{vmatrix} z+p_h\mu_h & -p_h\mu_h \\ -p_h\mu_h & z+p_h\mu_h \end{vmatrix} =$$

$$= (z+\lambda)z^2[(z+p_h\mu_h)^2 - (p_h\mu_h)^2] = (z+\lambda)z^3(z+2p_h\mu_h)$$

and that $\boldsymbol{m}_h^2 = \boldsymbol{m}_h^3 = \mathbf{0}$ and $\boldsymbol{m}_h^4 + \boldsymbol{m}_h^5 = -\frac{\lambda}{\mu_h}\boldsymbol{m}_h^1$, where $\boldsymbol{m}_h^i$ denotes the i-th column on M_h. The stated values are then the solution of $M_h\boldsymbol{x} = \mathbf{0}$ under the side conditions $C_0 + C_1 + C_2 = C$ and $S_1 + S_2 = S$. □

Here, the implication is that for any two queueing networks with the same mean service times but different (high-SCV) Coxian distributions, the steady-state queue lengths are identical under the HLR assumption. But, comparing this result against Theorem 1, we can say more: Regardless of whether the Coxian distribution is low-SCV or high-SCV, the fluid steady-state queue lengths are insensitive.

3.3 Exponential Distribution

The purpose of this section is to show that, if the Coxian-distributed queue is replaced by an exponential one with the same average service time $1/\mu_0$, then under the HLR assumption we find that the steady-state queue lengths will coincide with those described in the previous two sub-sections.

Let us recall that the ODE system for such a case is (5). The HLR assumption requires $C_1 \geq S_1$. Hence, we can write the resulting linear ODE system as

$$\dot{\boldsymbol{x}} = M_e\,\boldsymbol{x}, \qquad \text{where } M_e = \begin{pmatrix} -\lambda & 0 & \mu_0 \\ \lambda & 0 & -\mu_0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Then, the following holds.

THEOREM 3. *For any given $\lambda, S > 0$, there exists a $\tilde{C} > 0$ such that for all $C \geq \tilde{C}$ the fluid approximation (5) subject to $x(0) = C\mathbb{1}_{C_1} + S\mathbb{1}_S$ has the following steady state:*

$$C_0 = \frac{\mu_0}{\lambda} S, \qquad C_1 = C - \frac{\mu_0}{\lambda} S.$$

Crucially, this and Theorems 1 and 2 show that the fluid steady-state queue lengths are the same if the mean service times in the networks are equal, regardless of whether exponential or Coxian-distributed service times are considered.

PROOF. It is clear that we can apply the very same reasoning as in Proposition 2 and 3 if we are able to show that, apart from 0, all eigenvalues of M_e have negative real parts and that the algebraic and geometric multiplicity of 0 are the same. This, however, is obvious because $\wp_{M_e}(z) = (z+\lambda)z^2$ and $\mathbf{m}_e^2 = 0$, $\boldsymbol{m}_e^3 = -\frac{\lambda}{\mu_0}\boldsymbol{m}_e^1$. The stated values are then the solution of $M_e\boldsymbol{x} = \mathbf{0}$ under the side condition $C_0 + C_1 = C$. □

3.4 Discussion

Taken together, Theorems 1, 2, and 3 allow us to establish a connection between the steady-state fluid insensitivity discussed in this paper and a bottleneck analysis of our queueing network. In all cases, i.e., low and high SCV as well

as exponentially distributed service times, the population of jobs in the delay station, C_0, is given by $C_0 = (\mu_0/\lambda)S$, where S is the total population of servers.

This result could be obtained by assuming a stochastic network with unitary utilization at the bottleneck queue, and calculating the performance measures of such network by applying Little's law [9]. Indeed, under this assumption, the network throughput in the steady-state is $\mu_0 S$, whereas the average response time at the delay station is simply $1/\lambda$. Note, however, that in contrast to the *heavy traffic* where only the number of customers tends to infinity and the numbers of servers is fixed, the fluid approximation refers to models where the populations of customers *and* servers are increased. Consequently, it is not a priori clear, whether the utilization will achieve one, meaning that the problem cannot be tackled by readily applying Little's law.

4. NUMERICAL EXPERIMENTS

The theoretical investigation conducted in the previous section has allowed us to conclude that the fluid analysis may not distinguish between queueing networks that differ only for higher-order moments than the means of their service time distributions. On the other hand, in Section 2.3 we have found that such networks have in general different stochastic behavior since the average queue lengths do depend on the SCV. This begs the question which of the service-time distributions is best approximated by fluid analysis when this is interpreted as an approximation of the stochastic process with finite populations, by $\boldsymbol{X}_v(t) \approx v\boldsymbol{x}(t)$.

In this section we carry out a numerical investigation that aims at answering this question. We do so by measuring the approximation error between the fluid solution and the expected queue length computed by CTMC analysis on a number of models. In all cases, we fixed $\mu_0 = 1.0$. We compared the following three service-time distributions: Erlang-10 (SCV 0.01), exponential (SCV 1), and a high-SCV Coxian distribution with SCV 50; these will be denoted by ERL-10, EXP, and SCV 50, respectively. We considered three values of λ: 0.8, 1.0, and 1.2. These were chosen because they are representative of a region of the parameter space where the fluid approximation does not behave well at relatively low populations. Hence, appreciable error trends may be observed as a function of the initial population levels. For each value of λ, we considered ERL-10, EXP, and SCV 50 with increasing populations of servers and jobs, according to the scaling of Kurtz discussed in Section 2.2. We fixed an initial configuration with $C = 10$ jobs and $S = 5$ servers, i.e., corresponding to the population sizes for $v = 1$, and scaled up the populations by setting $v = 2, 3, 4, 5$. Hence, for instance, the case $v = 5$ corresponds to 50 jobs and 25 servers in the network. For each of these cases, we measured the approximation error as the absolute difference between the steady-state fluid queue length and the CTMC queue length at the delay station, divided by v: This is an error metric between the fluid model and the *density* CTMC process.

All the computations were performed in Matlab. We used the built-in `ode15s` function for the numerical integration of the ODEs. We checked convergence to the steady state by ensuring that the norm of the time-derivative of the ODE solution was less 1E-6. Furthermore, we verified that every ODE solution satisfied the HLR condition by checking that it held at every time-point returned by the numerical integrator. The CTMCs were solved by simulation using the method of batch means, terminating when the confidence interval at 99% confidence level was within 1% from the statistical average.

The results are reported in Figure 5, which plots these approximation errors for all cases considered. The most important observation is that, in all cases, the queueing networks with smaller SCV of the service-time distributions are closer to the fluid solution than those with higher SCV. This confirms the intuitive idea that higher SCVs introduce more noise into the system, making the fluid approximation less accurate for finite and relatively low populations. Additionally, the error trends confirm that the approximation tends to improve with v. Here, all errors fall below 5% as early as for $v = 5$. Finally, we observe that the errors depend on λ: the highest approximation errors were registered for $\lambda = 1.0$. In this example, using values outside the range $[0.8, 1.2]$ lead to excellent accuracy of the fluid approximation already for $v = 1$ and for all service-time distributions, making the error trends less informative than those reported here.

5. CONCLUSION

This paper has identified a class of queueing models which feature insensitivity of the fluid estimates of the steady-state queue lengths. We considered a Coxian distributed queue fed by an exponential delay station in tandem. Under the assumption of a heavy-load regime, where the fluid trajectories are such that in the Coxian queue there are consistently more jobs than available servers across the entire time horizon, we found that the steady-state queue lengths only depend on the average service times of the network.

This fact appears unsurprising when one interprets the fluid model as the deterministic asymptotic behavior of a (stochastic) sequence of queueing networks. In the limit of infinite numbers of jobs and servers, the stochastic noise will vanish and the system's dynamics will only be dependent on the average drift of the process, which is indeed the main observation of this paper. On the other hand, our result is surprising because insensitivity only occurs in the steady state, while the transient fluid behavior is in general affected by higher-order moments of the service-time distribution. Furthermore, we were able to find systems of differential equations that are inherently different — because they have different rate parameters and different sizes — yet they give essentially the same behavior.

It has to be pointed out that the interpretation of a fluid steady state is not obvious. The limit result by Kurtz used in this paper guarantees convergence in probability over a *finite* time interval, while we get insensitivity of the fluid trajectories when time goes to *infinity*. Convergence has been extended to the steady-state, but only under the condition of a unique globally attracting stationary point for the ODE (e.g., [6]). As discussed, our ODEs do not enjoy this property due to their piecewise linearity. However, this does not impinge on the usefulness of our result: Insensitivity in the steady state implies that there always exists a *finite time point* such that the difference between the queue lengths of any two tandem networks with the same average service times is less than any given threshold.

Finally, we comment on two aspects that may affect the generality of the results presented in this paper. The first aspect concerns the assumption on the heavy-load regime. It is only under this condition that we are able to formally prove insensitivity. However, we have found (in tests that

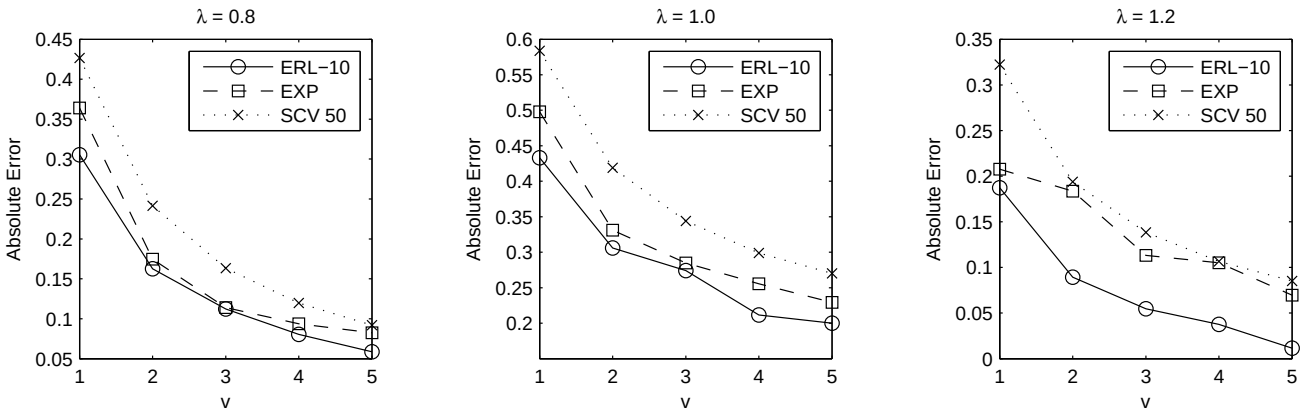

Figure 5: Approximation errors between the steady-state fluid estimate and the CTMC average of the queue length at the delay station for $\lambda = 0.8$, $\lambda = 1.0$, and $\lambda = 1.2$.

are not reported here) that insensitivity may hold in other regions of the ODE solution space. The second aspect concerns the fact that insensitivity was proven only for a specific model, our tandem queueing network of Figure 1. While it can be seen as prototypical high-level model of a system that is wholly abstracted away by a single generally distributed queue in a closed-world assumption, the proof techniques are not directly applicable to networks with arbitrary topologies.

It will be the subject of future work to address these two aforementioned open issues.

Acknowledgment

The authors wish to thank Giuliano Casale for his very insightful comments on earlier drafts of this paper. This work was partially supported by the DFG project FEMPA, TR 1120/1-1, and by the EU project QUANTICOL, 600708. Much of this work was done while the authors were at the Department for Informatics of Ludwig Maximilians University of Munich, Germany.

6. REFERENCES

[1] J. Anselmi and I. Verloop. Energy-aware capacity scaling in virtualized environments with performance guarantees. *Performance Evaluation*, 68(11):1207 – 1221, 2011. *Special Issue Performance 2011*.

[2] G. Bolch, S. Greiner, H. de Meer, and K. Trivedi. *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. Wiley, 2005.

[3] M. Bramson, Y. Lu, and B. Prabhakar. Randomized load balancing with general service time distributions. In V. Misra, P. Barford, and M. S. Squillante, editors, *SIGMETRICS*, pages 275–286. ACM, 2010.

[4] R. Cooper. *Introduction to Queueing Theory*. North-Holland, 1981.

[5] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

[6] N. Gast and B. Gaujal. A Mean Field Model of Work Stealing in Large-Scale Systems. *ACM SIGMETRICS*, pages 13–24, 2010.

[7] S. Gershgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Izvestija Akademii Nauk SSSR, Serija Matematika*, 7(3):749–754, 1931.

[8] T. G. Kurtz. Solutions of ordinary differential equations as limits of pure Markov processes. *J. Appl. Prob.*, 7(1):49–58, April 1970.

[9] J. Little. A Proof of the Queuing Formula: $L = \lambda W$. *Operations Research*, 9(3):383–387, 1961.

[10] M. Miyazawa. Insensitivity and product-form decomposability of reallocatable GSMP. *Advances in Applied Probability*, 25(2):415–437, 1993.

[11] R. Schassberger. Two remarks on insensitive stochastic models. *Advances in Applied Probability*, 18(3):791–814, 1986.

[12] G. Sharma, A. Ganesh, and P. Key. Performance analysis of contention based medium access control protocols. *IEEE Transactions on Information Theory*, 55(4):1665–1682, 2009.

[13] W. J. Stewart. *Probability, Markov Chains, Queues, and Simulation*. Princeton University Press, 2009.

[14] B. Van Houdt and L. Bortolussi. Fluid limit of an asynchronous optical packet switch with shared per link full range wavelength conversion. In P. G. Harrison, M. F. Arlitt, and G. Casale, editors, *SIGMETRICS*, pages 113–124. ACM, 2012.

[15] W. Walter and R. Thompson. *Ordinary Differential Equations*. Springer, 1998.

[16] W. Whitt. *Stochastic-process limits*. Springer, 2002.

[17] P. Whittle. Partial balance and insensitivity. *Journal of Applied Probability*, 22(1):168–176, 1985.

[18] X. Zhou, S. Ioannidis, and L. Massoulie. On the stability and optimality of universal swarms. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, SIGMETRICS '11, pages 341–352, New York, NY, USA, 2011. ACM.

APPENDIX

In this appendix we present the proofs of the Propositions stated in the paper.

Proof Proposition 1. We start our proof with a derivation of the characteristic polynomial $\wp_{M_l^n}(z) = |zI - M_l^n|$

of M_l^n, where $|\cdot|$ denotes the determinant and $z \in \mathbb{C}$. For this, we first observe that

$$\wp_{M_l^n}(z) = \left| zI - \left(\begin{array}{c|c} A_n & B_n \\ \hline 0 & D_n \end{array}\right)\right| = \wp_{A_n}(z) \cdot \wp_{D_n}(z)$$

and, thanks to Laplace's expansion theorem,

$$\begin{aligned}
\wp_{A_n}(z) &= (z+\lambda)\cdot z^n \\
\wp_{D_n}(z) &= (z+\mu_l p_l)(z+\mu_l)^{n-1} + \\
&\quad + (-1)^{n+1}(-\mu_l)(-\mu_l p_l)(-\mu_l)^{n-2} \\
&= (z+\mu_l p_l)(z+\mu_l)^{n-1} + (-1)^{2n}\mu_l^{n-1}(-\mu_l p_l) \\
&= (z+\mu_l p_l)(z+\mu_l)^{n-1} - \mu_l p_l^n.
\end{aligned}$$

Note that $\wp_{D_n}(0) = 0$ and $\dot{\wp}_{D_n}(0) \neq 0$, since

$$\dot{\wp}_{D_n}(z) = (z+\mu_l)^{n-1} + (n-1)(z+\mu_l p_l)(z+\mu_l)^{n-2}.$$

Thus, there exists a polynomial q such that $\wp_{D_n}(z) = z\cdot q(z)$ and $q(0) \neq 0$, which implies that the algebraic multiplicity of 0 is $n+1$, because

$$\wp_{M_l^n}(z) = \wp_{A_n}(z)\cdot\wp_{D_n}(z) = (z+\lambda)\cdot z^n \cdot z \cdot q(z).$$

Recall that the geometric multiplicity of an eigenvalue is always less or equal than the algebraic one. Thus, in order to see that the geometric multiplicity of 0 is $n+1$ as well, it is sufficient to show that the dimension of $\{\boldsymbol{x} \in \mathbb{R}^{2n+1} \mid M_l^n \boldsymbol{x} = \boldsymbol{x}\}$ is at least $n+1$. This follows, however, by observing that the columns $\boldsymbol{m}_l^2, \boldsymbol{m}_l^3, \ldots, \boldsymbol{m}_l^{n+1}$ of M_l^n are zero valued and that

$$\begin{aligned}
\boldsymbol{m}_l^{n+2} + p_l\boldsymbol{m}_l^{n+3} + \ldots + p_l\boldsymbol{m}_l^{2n+1} = \\
= (\mu_l, -\mu_l, 0, \ldots, 0)^T = -\frac{\mu_l}{\lambda}\boldsymbol{m}_l^1.
\end{aligned}$$

The second claim follows if one can show that, apart from 0, all roots of $\wp_{D_n}(z)$ have negative real parts. For this, let us denote the rows of D_n^T by $d_1, \ldots, d_n \in \mathbb{R}^n$. The statement follows then from $\wp_{D_n}(z) = \wp_{D_n^T}(z)$, Gershgorin's circle theorem [7] and

$$-d_{ii} = \sum_{\substack{1\leq j\leq n,\\ j\neq i}} |d_{ij}|$$

for all $1 \leq i \leq n$. □

Proof Proposition 2. Without special mentioning, we apply in the following results from the area of systems of linear differential equations with constant coefficients. For a detailed discussion on the latter, see [15]. Let $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{2n+1}$ be the basis of the solution set of (6). Using Proposition 1, we may assume without loss of generality that $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{n+1}$ are induced by the eigenvalue 0, whereas $\boldsymbol{\Theta}_{n+2}, \ldots, \boldsymbol{\Theta}_{2n+1}$ are induced by the remaining eigenvalues which have all a negative real part. Since the geometric and algebraic multiplicity of 0 coincide, there are $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{n+1} \in \mathbb{R}^{2n+1}$ such that $\boldsymbol{\Theta}_i(t) = \boldsymbol{\xi}_i$ for all $t \geq 0$ and $1 \leq i \leq n+1$. Let $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ denote the solution induced by $\boldsymbol{\phi}(0) = C\mathbb{1}_{C_1}$ and $\boldsymbol{\psi}(0) = S\mathbb{1}_{S_1}$, respectively. Then

$$\begin{aligned}
\boldsymbol{\phi}(t) &= C\sum_{i=1}^{2n+1}\alpha_i\boldsymbol{\Theta}_i(t) = C\Big(\sum_{i=1}^{n+1}\alpha_i\boldsymbol{\xi}_i + \sum_{i=n+2}^{2n+1}\alpha_i\boldsymbol{\Theta}_i(t)\Big) \\
&= C\boldsymbol{\phi}^0 + C\boldsymbol{\phi}^-(t) \\
\boldsymbol{\psi}(t) &= S\sum_{i=1}^{2n+1}\beta_i\boldsymbol{\Theta}_i(t) = S\Big(\sum_{i=1}^{n+1}\beta_i\boldsymbol{\xi}_i + \sum_{i=n+2}^{2n+1}\beta_i\boldsymbol{\Theta}_i(t)\Big) \\
&= S\boldsymbol{\psi}^0 + S\boldsymbol{\psi}^-(t)
\end{aligned}$$

Note that $\boldsymbol{x} \equiv \boldsymbol{\phi} + \boldsymbol{\psi}$ thanks to the linearity of the system and the uniqueness of the solution. This motivates to define $\boldsymbol{x}_{ss} := C\boldsymbol{\phi}^0 + S\boldsymbol{\psi}^0$ and $\boldsymbol{x}_{tr}(t) := \boldsymbol{x}(t) - \boldsymbol{x}_{ss}$. Together with $\wp_{M_l^n}(z) := |zI - M_l^n|$ and

$$a := |\max\{\Re(z) \mid \wp_{M_l^n}(z) = 0 \wedge z \neq 0\}|,$$

it then holds that

$$\begin{aligned}
\|\boldsymbol{\phi}^-(t) + \boldsymbol{\psi}^-(t)\|_\infty &\leq \|\boldsymbol{\phi}^-(t)\|_\infty + \|\boldsymbol{\psi}^-(t)\|_\infty \\
&\leq \sum_{i=n+2}^{2n+1}|\alpha_i|e^{-at}(1+t)^n + \sum_{i=n+2}^{2n+1}|\beta_i|e^{-at}(1+t)^n \\
&= b_1 e^{-at}(1+t)^n + b_2 e^{-at}(1+t)^n,
\end{aligned}$$

which implies

$$\|\boldsymbol{x}_{tr}(t)\|_\infty \leq (b_1 C + b_2 S)e^{-at}(1+t)^n.$$

□

Proof Proposition 3. Since $S_1, \ldots, S_n$ and $C_2, \ldots, C_n$ do not depend on C_0, C_1 in (6) and $C_i(0) = S_i(0)$, $\dot{C}_i = \dot{S}_i$ for all $2 \leq i \leq n$, we are left with the case $i = 1$. By Proposition 2, we know that $\boldsymbol{x}(t) = \boldsymbol{x}_{ss} + \boldsymbol{x}_{tr}(t)$ with

$$\|\boldsymbol{x}_{tr}(t)\|_\infty \leq (b_1 C + b_2 S)e^{-at}(1+t)^n, \quad \text{for all} \quad t \geq 0.$$

Using $J_C := (b_1 C + b_2 S)$, we infer that

$$J_C e^{-at}(1+t)^n \leq 1 \Leftrightarrow t - \frac{n}{a}\log(1+t) \geq -\frac{1}{a}\log\left(\frac{1}{J_C}\right),$$

meaning that $\log(1+t) \leq \frac{a}{2n}t$ and $t \geq -\frac{2}{a}\log(\frac{1}{J_C})$ imply $J_C e^{-at}(1+t)^n \leq 1$. Note that $t \mapsto J_C e^{-at}(1+t)^n$ is decreasing for $t > \frac{n}{a} - 1$. Hence, since $t_0 := -\frac{2}{a}\log(\frac{1}{J_C})$ satisfies $\log(1+t_0) \leq \frac{a}{2n}t_0$ and $t_0 > \frac{n}{a} - 1$ if $C > 0$ is large enough, it suffices to prove that there exists a $\tilde{C} > 0$ such that

$$C \geq S\left(1 - \frac{2}{a}\log\left(\frac{1}{b_1 C + b_2 S}\right)\right) + 2, \quad \text{if} \quad C \geq \tilde{C},$$

because of

$$C_1(t_0) \geq C - St_0 = C + \frac{2S}{a}\log\left(\frac{1}{J_C}\right) \geq S + 2 \geq S_1(t_0).$$

This, however, follows from L'Hospital's theorem which ensures that

$$\lim_{C\to\infty}\frac{S\left(1 - \frac{2}{a}\log\left(\frac{1}{b_1 C + b_2 S}\right)\right) + 2}{C} = 0$$

if and only if $\lim_{C\to\infty}\frac{2S}{a}\frac{b_1}{b_1 C + b_2 S} = 0$. □