Analysis of a two-layered network with correlated queues by means of the power-series algorithm

EURANDOM PREPRINT SERIES
2012-005

April 5, 2012

Analysis of a two-layered network with correlated queues by means of the power-series algorithm

J.L. Dorsman, R.D. van der Mei, M. Vlasiou
ISSN 1389-2355

Analysis of a two-layered network with correlated queues by means of the power-series algorithm

J.L. Dorsman [*] [†]
j.l.dorsman@tue.nl

R.D. van der Mei [†] [‡]
R.D.van.der.Mei@cwi.nl

M. Vlasiou [*] [†]
m.vlasiou@tue.nl

April 5, 2012

Abstract

We consider an extension of the classical machine-repair model, also known as the computer-terminal model or time-sharing model. As opposed to the classical model, we assume that the machines, apart from receiving service from the repairman, supply service themselves to queues of products. The extended model can be viewed as a two-layered queueing network, of which the first layer consists of two separate queues of products. Each of these queues is served by its own machine. The marginal and joint queue length distributions of the first-layer queues are hard to analyse in an exact fashion. Therefore, we apply the power-series algorithm to this model to obtain the light-traffic behaviour of the queue lengths symbolically. This leads to two accurate approximations for the marginal mean queue length. The first approximation, based on the light-traffic behaviour, is in closed form. The second approximation is based on an interpolation between the light-traffic behaviour and heavy-traffic results for the mean queue length. The obtained approximations are shown to work well for arbitrary loaded systems. The proposed numerical algorithm and approximations may prove to be very useful for system design and optimisation purposes in application areas such as manufacturing, computer systems and telecommunications.

1 Introduction

In this paper, we study a layered queueing network (LQN) consisting of two layers. We define an LQN to be a queueing network where in addition to the traditional "servers" and "customers", there exist customer units that act as servers for upper-layer customers. Thus, the network can be decomposed into multiple layers, in each of which units act as either strictly a customer or a server. Layered queueing networks occur naturally in information and e-commerce systems, grid systems, and real-time systems such as telecom switches, see [10] and references therein for an overview.

The LQN under consideration is motivated by a two-fold extension of the traditional machine-repair model. This model, also known as the *computer terminal model* (cf. [2]) or as the *time sharing system* (cf. [16, Section 4.11]), is a well-studied problem in the literature. In the machine-repair model, there is a number of machines (two in our case) working in parallel, and one repairman. As soon as a machine fails, it joins a repair queue in order to be repaired by the repairman. It is one of the key models to describe problems with a finite input population. A fairly extensive analysis of the machine-repair model can be found in Takács [19, Chapter 5].

Funded in the framework of the STAR-project "Multilayered queueing systems" by the Netherlands Organization for Scientific Research (NWO). The research of M. Vlasiou is also partly supported by an NWO individual grant through project 632.003.002. The work of the second author has been carried out in the context of the IOP GenCom project Service Optimization and Quality (SeQual), which is supported by the Dutch Ministry of Economic Affairs, Agriculture and Innovation via its agency Agentschap NL.

[*] EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[†] Probability and Stochastic Networks, Centum Wiskunde & Informatica (CWI), P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

[‡] Department of Mathematics, VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

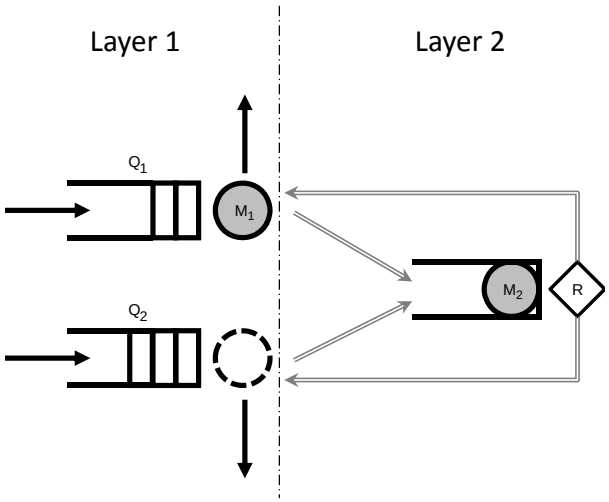

Figure 1: The two-layered model under consideration.

We extend this model in two directions. First, we allow different machines to have mutually different uptime or repair time distributions. As observed in [12], this leads to technical complications. For example, the arrival theorem (cf. [17]) cannot be used any more to derive the stationary downtime distribution of the machines as is done for the original model in [21]. Secondly, we assume that each of the machines processes a stream of products, which leads to the addition of queues in front of the machines. Observe that in this case a machine has a dual role. As in the traditional model, the machine has a *customer role* with respect to the repairman, but it now also has a *server role* with respect to the products. This leads to the formulation of a LQN with two layers, which we also refer to as the *two-layered model* or simply the *layered model*. This extension has immediate applications in manufacturing, but is also of interest for other application areas, such as telecommunication systems. For instance, this extension of the machine-repair model occurs naturally in the modelling of middleware technology, where multi-threaded application servers compete for access to shared object code. Access of threads to the object code is typically handled by a portable object adapter (POA) that serialises the threads trying to access a shared object. During the complete duration of the serialisation and execution time a thread typically remains blocked, and upon finalizing the execution, the thread is de-activated and ready to process pending requests [13]. In this setting, the application servers and the shared object code are analogous to the machines and the repairman respectively in the machine-repair model.

Although applications of the two-layered model are not restricted to manufacturing, we refer to the entities in the model as *products*, *machines* and the *repairman* respectively. The first layer of the resulting LQN contains two queues of products, see Figure 1. Each of these queues is served by its own machine. At any point in time, a machine is subject to breakdowns, irrespective of the state of the first-layer queue. When a machine breaks down, the service of a product in progress is aborted and starts anew once the machine becomes operational again. For ease of discussion, we assume that there are two machines only, as opposed to the classical machine-repair model. As will be evident in the sequel, the approach we follow can be readily extended to more machines or repairmen, but certain computations become increasingly cumbersome. The second layer consists of a repairman and a repair buffer. If, upon breakdown of a machine, the repairman is idle, the machine is immediately taken into service. Once the machine is again operational, it starts serving products once more. When the repairman is busy repairing another machine however, the machine waits in the buffer. As soon as the second machine is repaired, the repair of the current machine starts.

An important feature of both the classical machine-repair model and the two-layered model under consideration is the fact that machines compete for repair facilities. As concluded in [8], this introduces significant positive dependencies in their downtimes and thus in the lengths of the queues in the first layer. The dependence between these queues makes exact analysis of their queue length distribution difficult. The amount of work present in a first-layer queue can be modelled as a reflected Markov modulated Lévy process, but its distribution is not easily derived

from that. Numerical evaluation, e.g. by simulation, may also be challenging. Especially when the model involves breakdowns and repairs occurring on a larger time scale than actual product arrivals and services, the computation time needed to achieve accurate results may be unacceptably long. Moreover, numerical methods are typically not transparent and provide little insight into parameter effects. Therefore, there is a need for symbolic, accurate and transparent queue length approximations which are easy to implement and are suitable for optimisation purposes.

In this paper, we derive two approximations for the mean queue length by applying the power-series algorithm (PSA) on the two-layered model. PSA is used to compute the steady-state distribution of multiple-queue systems, which fit in the class of multi-dimensional quasi birth-and-death processes (QBDs). The basic idea of this algorithm stems from Hooghiemstra et al. [14]. The algorithm has been further developed by Blanc (see e.g. [3, 4]). For an overview of PSA, as well as the initial literature on PSA, see [5]. The use of PSA is, in many regards, advantageous over numerical methods such as simulation. The computation time needed to achieve accurate numerical results is generally much less, especially for lightly loaded systems. More importantly, the computational scheme provided by PSA can also be executed symbolically to obtain closed-form expressions for moments and cross-moments of the queue length distributions, in theory up to arbitrary precision. In practice, symbolically only the light-traffic (LT) behaviour can be computed. Based on this behaviour, we obtain two approximations for the mean queue length, which is the main result of this paper.

In Section 2, the two-layered model is described explicitly and the notation required is given. Then, we explain how to implement PSA for this model in Section 3. The resulting symbolic expressions for the light-traffic behaviour are described in Section 4. Based on the light-traffic behaviour, we propose two approximations for the mean queue length in Section 5. The first one, as provided in Section 5.1, is in closed form. It is therefore suitable for implementation and optimisation purposes. Moreover, the approximation is exact in light traffic and performs well for general loaded systems. To improve the accuracy of the approximation even further, we provide a second approximation in Section 5.2 that is also exact when the queue is fully saturated. This approximation is based on an interpolation between the light-traffic behaviour and heavy-traffic (HT) results. This approximation is very accurate. It competes with the numerical errors made by simulation or PSA. Moreover, the two approximations greatly illustrate the effects of the model parameters onto the queue length.

2 Model description and notation

The model considered in this paper consists of two layers, see Figure 1. The first layer consists of two machines M_1 and M_2 as well as the corresponding queues Q_1 and Q_2, which we will refer to as first-layer queues. Products arrive at Q_i according to a Poisson process with rate λ_i. The service requirement of a product in Q_i is exponentially distributed with parameter μ_i. We denote the load offered to Q_i by $\rho_i = \frac{\lambda_i}{\mu_i}$. The steady-state queue length of Q_i, including the product in service, is denoted by L_i. The delay, or the waiting time, incurred by a type-i product before it enters service is denoted by D_i. Furthermore, the time between the arrival of a type-i product and the end of its service is referred to as the sojourn time S_i. After an exponentially (σ_i) distributed uptime or lifetime, denoted by U_i, the machine M_i serving Q_i will break down, and the service of Q_i stops. The service of a product in progress is then aborted, and will be resumed once the machine is operational again. When a machine breaks down, it moves to the repair queue, where it will wait if the repairman is busy repairing the other machine; otherwise the repair will start immediately. Thus, a downtime of a machine consists of a repair time and possibly a waiting time. The time needed for a repairman to return M_i to an operational state is exponentially (ν_i) distributed. After a repair, the machine returns to Q_i and commences service again.

In various computations throughout this paper, we need to keep track of the background environment, namely whether the two machines are working or not. To this end, let $\{\mathbf{\Phi}(t), t \geq 0\}$ be the Markov process describing the state of the machines M_1 and M_2. More specifically, $\mathbf{\Phi}(t) = \{\Phi_1(t), \Phi_2(t)\}$ specifies for each machine whether it is up (U), in repair (R) or waiting for repair (W) at time t. This Markov process operates on the state space $\mathcal{S} := \{(U,U), (U,R), (R,U), (R,W), (W,R)\}$ with generator matrix $Q^{\mathbf{\Phi}}$. Its stationary distribution vector $\boldsymbol{\pi}^{\mathbf{\Phi}}$ is uniquely determined by the equations $\boldsymbol{\pi}^{\mathbf{\Phi}} Q^{\mathbf{\Phi}} = \mathbf{0}$ and $\sum_{j \in \mathcal{S}} \boldsymbol{\pi}^{\mathbf{\Phi}}_j = 1$.

The queue length of a first-layer queue depends heavily on the availability of its machine in the past. To keep track

of the latter, let $C(t)$ represent the amount of time M_1 has been in an up state in the time period $[0, t)$. Assuming the process $\{\mathbf{\Phi}(t), t \geq 0\}$ is already in stationarity at $t = 0$, $C(t)$ is defined as

$$C(t) = \int_{s=0}^{t} \mathbb{1}_{\{\Phi_1(s)=U\}} ds. \tag{1}$$

The long-run time-averaged mean of the process $\{C(t), t \geq 0\}$ is given by

$$m_C := \lim_{t\to\infty} \frac{\mathbb{E}[C(t)]}{t} = \lim_{t\to\infty} \frac{\int_{s=0}^{t} \mathbb{P}(\Phi_1(s) = U)ds}{t} = \boldsymbol{\pi}^{\mathbf{\Phi}}_{(U,U)} + \boldsymbol{\pi}^{\mathbf{\Phi}}_{(U,R)}.$$

Note that by standard renewal arguments we also have that $m_C = \frac{\mathbb{E}[U_1]}{\mathbb{E}[U_1]+\mathbb{E}[D_1]}$. Furthermore, the long-run time-averaged variance is given by

$$\sigma_C^2 := \lim_{t\to\infty} \frac{\text{Var}[C(t)]}{t}.$$

To determine $\text{Var}[C(t)]$, view $\{(C(t), \mathbf{\Phi}(t)), t \geq 0\}$ as a Markov additive process (MAP) with matrix exponent $F(s) = \text{diag}(-s, -s, 0, 0, 0) + Q^{\mathbf{\Phi}}$. By standard theory on MAPs (see [1, p. 311–312], the matrix $Z(s, t)$ with elements $Z_{i,j}(s, t) = \mathbb{E}[e^{-sC(t)}\mathbb{1}_{\{\mathbf{\Phi}(t)=j\}}|\mathbf{\Phi}(0) = i] = \mathbb{P}(\mathbf{\Phi}(t) = j|\mathbf{\Phi}(0) = i)\mathbb{E}[e^{-sC(t)}|\mathbf{\Phi}(0) = i, \mathbf{\Phi}(t) = j]$, is determined by

$$Z(s, t) = e^{tF(s)}. \tag{2}$$

Note that $\mathbb{P}(\mathbf{\Phi}(t) = j|\mathbf{\Phi}(0) = i)$ is obtained by computing $Z_{i,j}(0, t)$. By conditioning, the Laplace-Stieltjes transform (LST) $\mathbb{E}[e^{-sC(t)}]$ is consequently easily derived from $Z(s, t)$, out of which $\text{Var}[C(t)]$ follows by differentiation with respect to s.

To keep track of the level of saturation of Q_1, we introduce the notion of normalised load. If M_1 never breaks down, then the stability condition for Q_1 reads $\rho_1 < 1$. However, in the case of breakdowns, this condition is not sufficient any longer, as M_1 only works for a fraction m_C of the time. We therefore define $\hat{\rho} := \frac{\rho_1}{m_C}$. We also refer to $\hat{\rho}$ as the normalised load of Q_1. Taking the breakdowns of M_1 into account, the stability condition for Q_1 is $\hat{\rho} < 1$.

Throughout this paper, we denote the L^1-norm of a vector $\boldsymbol{v}$ consisting of n elements by $|\boldsymbol{v}| = v_1 + \cdots + v_n$. The vector $\mathbf{0} \in \mathbb{N}^n$ represents the n-th dimensional vector which consists only of zeros. The vector $\boldsymbol{e}_j \in \mathbb{N}^n$ represents the unit vector of which the j-th entry equals one. We denote the indicator function on the event A by $\mathbb{1}_{\{A\}}$. Finally, for two functions $f(x)$ and $g(x)$, we write $f(x) = \mathcal{O}(g(x))$ if $\lim_{x\downarrow 0} |f(x)/g(x)| < \infty$.

3 Application of the Power-Series Algorithm

In this section, we show how PSA can be used to analyse the two-layered model. PSA is typically used to compute the steady-state distribution of multiple-queue systems, which fit in the class of quasi birth-and-death processes (QBDs). The two-layered model is such a multi-dimensional QBD and consists of two components. The first component, $\boldsymbol{L}(t) = \{L_1(t), L_2(t)\}$, describes the queue length at each of the queues. The second component models any non-exponentiality in the system. In our system, non-exponentiality is caused by the fact that the machines alternate between up-times and downtimes and is represented by $\mathbf{\Phi}(t)$. Thus, $\{(\boldsymbol{L}(t), \mathbf{\Phi}(t)), t \geq 0\}$ can be seen as a Markov process on the state space $\mathbb{N}^2 \times \mathcal{S}$. When the system is stable, the steady-state probabilities $p(\boldsymbol{l}, \boldsymbol{\varphi}), (\boldsymbol{l}, \boldsymbol{\varphi}) \in \mathbb{N}^2 \times \mathcal{S}$ can be obtained in principle by solving the set of global balance equations. However, due to the multi-dimensionality of this process, this set of equations cannot be solved recursively, and is therefore hard to solve in practice. The intrinsic idea behind PSA is the transformation of the non-recursively solvable set of balance equations into a recursively solvable set of equations by adding one dimension into the state space. This is achieved by expressing the steady-state probabilities as power series in some variable based on the model parameters, and allows practical calculation of steady-state probabilities. As a result, performance measures of the form $\mathbb{E}[g(\boldsymbol{L}, \mathbf{\Phi})]$ can be computed, where $g(\cdot)$ is an arbitrary function, and $(\boldsymbol{L}, \mathbf{\Phi}) = \lim_{t\to\infty}(\boldsymbol{L}(t), \mathbf{\Phi}(t))$. We first define the one-step transition rates and the global balance equations corresponding to the Markov process $\{(\boldsymbol{L}(t), \mathbf{\Phi}(t)), t \geq 0\}$ in Section 3.1. Then, we apply PSA directly to the two-layered model in Section 3.2.

3.1 Preliminaries

Before applying PSA, we study the Markov Process $\{(\boldsymbol{L}(t), \boldsymbol{\Phi}(t)), t \geq 0\}$ and consider its one-step transition rates and global balance equations.

The one-step transition rate corresponding to the transition from state $(\boldsymbol{l}, \boldsymbol{\varphi}) \in \mathbb{N}^2 \times \mathcal{S}$ to state $(\boldsymbol{l} + \boldsymbol{e}_i, \boldsymbol{\varphi})$ equals the arrival rate λ_i. However, in order to fully exploit the flexibility that PSA provides, we specify each of the arrival rates by a "relative" arrival rate $a^{(i)}(\boldsymbol{l}, \boldsymbol{\varphi})$ times a constant χ. The quantity χ will be used by PSA to introduce another dimension to the state space. For $(\boldsymbol{l}, \boldsymbol{\varphi}) \in \mathbb{N}^2 \times \mathcal{S}$ and $\boldsymbol{\psi} \in \mathcal{S}$, we define the one-step transition rates as follows:

$\chi a^{(j)}(\boldsymbol{l}, \boldsymbol{\varphi})$:	the arrival rate at Q_j at state $(\boldsymbol{l}, \boldsymbol{\varphi})$, leading to a transition to state $(\boldsymbol{l} + \boldsymbol{e}_j, \boldsymbol{\varphi})$, $j = 1, 2$,
$d^{(j)}(\boldsymbol{l}, \boldsymbol{\varphi})$:	the departure rate from Q_j at state $(\boldsymbol{l}, \boldsymbol{\varphi})$, leading to a transition to state $(\boldsymbol{l} - \boldsymbol{e}_j, \boldsymbol{\varphi})$, with $d^{(j)}(\boldsymbol{l}, \boldsymbol{\varphi}) = 0$ if $n_j = 0, j = 1, 2$,
$u(\boldsymbol{l}, \boldsymbol{\varphi}, \boldsymbol{\psi})$:	the transition rate from $(\boldsymbol{l}, \boldsymbol{\varphi})$ to $(\boldsymbol{l}, \boldsymbol{\psi})$.

Linking this with the notation given in Section 2, this means that for $(\boldsymbol{l}, \boldsymbol{\varphi}) \in \mathbb{N}^2 \times \mathcal{S}$ and $j = 1, 2$:

$$
\begin{aligned}
&\chi a^{(j)}(\boldsymbol{l}, \boldsymbol{\varphi}) = \lambda_j, \\
&d^{(j)}(\boldsymbol{l}, \boldsymbol{\varphi}) = \mu_j \mathbb{1}_{\{n_j > 0\}} \mathbb{1}_{\{\varphi_j = U\}}, \\
&u(\boldsymbol{l}, (U,U), (R,U)) = u(\boldsymbol{l}, (U,R), (W,R)) = \sigma_1, \\
&u(\boldsymbol{l}, (U,U), (U,R)) = u(\boldsymbol{l}, (R,U), (R,W)) = \sigma_2, \\
&u(\boldsymbol{l}, (R,U), (U,U)) = u(\boldsymbol{l}, (R,W), (U,R)) = \nu_1, \\
&u(\boldsymbol{l}, (U,R), (U,U)) = u(\boldsymbol{l}, (W,R), (R,U)) = \nu_2.
\end{aligned}
$$

It remains to choose an appropriate value for χ. For the application of PSA, it is generally required that there exists a positive real χ^* such that both Q_1 and Q_2 are stable for $0 \leq \chi < \chi^*$. To satisfy this requirement, we choose

$$\chi = \hat{\rho} = \frac{\lambda_1}{\mu_1 m_C}. \tag{3}$$

This leads to $a^{(1)}(\boldsymbol{l}, \boldsymbol{\varphi}) = \mu_1 m_C$ and $a^{(2)}(\boldsymbol{l}, \boldsymbol{\varphi}) = \frac{\lambda_2}{\lambda_1} \mu_1 m_C$. Note that for the current choice of χ, there indeed exists an upper bound below which both queues are stable. Evidently, when the normalised workload does not exceed one, Q_1 is stable. Moreover, the ratio between μ_1 and μ_2, as well as the ratio between the fraction of time M_2 is up and m_C are assumed to be finite; i.e., we assume that none of the service rates and time fractions are zero. Thus, there must exist a positive real c, such that Q_2 is stable whenever $0 \leq \chi \leq c$. As a result, the requirement is satisfied when taking $\chi^* = \max\{1, c\}$.

The global balance equations of the Markov process $\{(\boldsymbol{L}(t), \boldsymbol{\Phi}(t)), t \geq 0\}$, expressed in the steady-state probabilities $p(\boldsymbol{l}, \boldsymbol{\varphi})$, are as follows:

$$
\begin{aligned}
&\left(\sum_{j=1}^{2} [\chi a^{(j)}(\boldsymbol{l}, \boldsymbol{\varphi}) + d^{(j)}(\boldsymbol{l}, \boldsymbol{\varphi})] + \sum_{\boldsymbol{\psi} \in \mathcal{S}} u(\boldsymbol{l}, \boldsymbol{\varphi}, \boldsymbol{\psi}) \right) p(\boldsymbol{l}, \boldsymbol{\varphi}) = \\
&\chi \sum_{j=1}^{2} a^{(j)}(\boldsymbol{l} - \boldsymbol{e}_j, \boldsymbol{\varphi}) p(\boldsymbol{l} - \boldsymbol{e}_j, \boldsymbol{\varphi}) \mathbb{1}_{\{\boldsymbol{l}_j > 0\}} + \sum_{j=1}^{2} d^{(j)}(\boldsymbol{l} + \boldsymbol{e}_j, \boldsymbol{\varphi}) p(\boldsymbol{l} + \boldsymbol{e}_j, \boldsymbol{\varphi}) \\
&+ \sum_{\boldsymbol{\psi} \in \mathcal{S}} u(\boldsymbol{l}, \boldsymbol{\psi}, \boldsymbol{\varphi}) p(\boldsymbol{l}, \boldsymbol{\psi}), \qquad (\boldsymbol{l}, \boldsymbol{\varphi}) \in \mathbb{N}^2 \times \mathcal{S}.
\end{aligned} \tag{4}
$$

We also have the normalisation equation

$$\sum_{(\boldsymbol{l}, \boldsymbol{\varphi}) \in \mathbb{N}^2 \times S} p(\boldsymbol{l}, \boldsymbol{\varphi}) = 1. \tag{5}$$

To substitute the steady-state probabilities, we observe the following property.

Property 3.1. For each state $(\boldsymbol{l}, \boldsymbol{\varphi})$, it holds that $p(\boldsymbol{l}, \boldsymbol{\varphi}) = \mathcal{O}(\chi^{|\boldsymbol{l}|})$. As is illustrated in [20], this property is valid for any QBD, where for each state $(\boldsymbol{l}, \boldsymbol{\varphi})$ with $\boldsymbol{l} \neq \boldsymbol{0}$, either $p(\boldsymbol{l}, \boldsymbol{\varphi}) = 0$, or there exists a path $\boldsymbol{\varphi}^{(0)}, \boldsymbol{\varphi}^{(1)}, \ldots, \boldsymbol{\varphi}^{(\nu)}$ in $\mathcal{S}$ for some ν, $0 \leq \nu < |\mathcal{S}|$, such that

$$\boldsymbol{\varphi}^{(0)} = \boldsymbol{\varphi}, u(\boldsymbol{l}, \boldsymbol{\varphi}^{(i-1)}, \boldsymbol{\varphi}^{(i)}) > 0, i = 1, \ldots, \nu,$$

and there is at least one queue with a non-zero departure rate in the state $(\boldsymbol{l}, \boldsymbol{\varphi}^{(\nu)})$. This condition is obviously met here, since there always exists such a path from any $\boldsymbol{\varphi} \in \mathcal{S}$ to the auxiliary state (U, U). In this state, both machines are up and departure rates for both of the queues are non-zero.

Based on Property 3.1, we introduce the following power series substitution for the steady-state probabilities:

$$p(\boldsymbol{l}, \boldsymbol{\varphi}) = \chi^{|\boldsymbol{l}|} \sum_{k=0}^{\infty} \chi^k b(k; \boldsymbol{l}, \boldsymbol{\varphi}), \qquad (\boldsymbol{l}, \boldsymbol{\varphi}) \in \mathbb{N}^2 \times \mathcal{S}. \tag{6}$$

3.2 Computational Scheme

In this section, we apply PSA to the two-layered model and derive a recursive, computational scheme for it. We obtain and solve a recursive set of equations for the coefficients $b(k; \boldsymbol{l}, \boldsymbol{\varphi})$ in (6). From this, all steady-state probabilities can be computed as well as any performance measures derived from them. We first substitute the power series expansion (6) into the balance equations (4). This leads to a polynomial expression in χ for both sides of the equations. By equating corresponding powers of χ, we obtain a recursion in the coefficients $b(k; \boldsymbol{l}, \boldsymbol{\varphi})$ for $k \in \mathbb{N}, (\boldsymbol{l}, \boldsymbol{\varphi}) \in \mathbb{N}^2 \times \mathcal{S}$. As a result, we can compute many performance measures by writing them as a power series in χ with different coefficients, but still involving the obtained values for $b(k; \boldsymbol{l}, \boldsymbol{\varphi})$ for $k \in \mathbb{N}, (\boldsymbol{l}, \boldsymbol{\varphi}) \in \mathbb{N}^2 \times \mathcal{S}$. Numerical computation of the performance measures is possible up to arbitrary precision, by starting the computational scheme with numerical values for the system parameters and applying the recursion until the desired accuracy is achieved. In theory, the performance measures can also be computed in a symbolic fashion. In practice however, only coefficients of very small order can be computed symbolically. Solving the recursive scheme becomes increasingly hard and the expressions involved become prohibitively complex as the corresponding order of χ increases.

We start by substituting the power-series expansion (6) into the balance equations (4). This implies the following set of equations for the coefficients $b(k; \boldsymbol{l}, \boldsymbol{\varphi})$:

$$\begin{aligned}
&\chi^{|\boldsymbol{l}|} \sum_{k=0}^{\infty} \chi^k \left(\sum_{j=1}^{2} [\chi a^{(j)}(\boldsymbol{l}, \boldsymbol{\varphi}) + d^{(j)}(\boldsymbol{l}, \boldsymbol{\varphi})] + \sum_{\boldsymbol{\psi} \in \mathcal{S}} u(\boldsymbol{l}, \boldsymbol{\varphi}, \boldsymbol{\psi}) \right) b(k; \boldsymbol{l}, \boldsymbol{\varphi}) = \\
&\chi^{|\boldsymbol{l}|-1} \sum_{k=0}^{\infty} \chi^k \sum_{j=1}^{2} \chi a^{(j)}(\boldsymbol{l} - \boldsymbol{e}_j, \boldsymbol{\varphi}) b(k; \boldsymbol{l} - \boldsymbol{e}_j, \boldsymbol{\varphi}) \mathbb{1}_{\{\boldsymbol{l}_j > 0\}} \\
&+ \chi^{|\boldsymbol{l}|+1} \sum_{k=0}^{\infty} \chi^k \sum_{j=1}^{2} d^{(j)}(\boldsymbol{l} + \boldsymbol{e}_j, \boldsymbol{\varphi}) b(k; \boldsymbol{l} + \boldsymbol{e}_j, \boldsymbol{\varphi}) \\
&+ \chi^{|\boldsymbol{l}|} \sum_{k=0}^{\infty} \chi^k \sum_{\boldsymbol{\psi} \in \mathcal{S}} u(\boldsymbol{l}, \boldsymbol{\psi}, \boldsymbol{\varphi}) b(k; \boldsymbol{l}, \boldsymbol{\psi}), \qquad (\boldsymbol{l}, \boldsymbol{\varphi}) \in \mathbb{N}^2 \times \mathcal{S}.
\end{aligned}$$

After eliminating the factor $\chi^{|\boldsymbol{l}|}$ from both sides of this set of equations, we obtain a polynomial equation of the form $\sum_{i=0}^{\infty} c_i \chi^i = \sum_{j=0}^{\infty} c_j \chi^j$. Since this equation holds for every $\chi \in [0, \chi^*)$, the coefficients of corresponding

powers of χ are equal. Thus, we have that $c_i = c_j$ for all $i = j$:

$$\Big(\sum_{j=1}^{2} d^{(j)}(\boldsymbol{l},\boldsymbol{\varphi}) + \sum_{\boldsymbol{\psi}\in\mathcal{S}} u(\boldsymbol{l},\boldsymbol{\varphi},\boldsymbol{\psi})\Big) b(k;\boldsymbol{l},\boldsymbol{\varphi}) =$$
$$\sum_{j=1}^{2} a^{(j)}(\boldsymbol{l}-\boldsymbol{e}_j,\boldsymbol{\varphi}) b(k;\boldsymbol{l}-\boldsymbol{e}_j,\boldsymbol{\varphi}) \mathbb{1}_{\{\boldsymbol{l}_j>0\}} - \sum_{j=1}^{2} a^{(j)}(\boldsymbol{l},\boldsymbol{\varphi}) b(k-1;\boldsymbol{l},\boldsymbol{\varphi}) \mathbb{1}_{\{k>0\}}$$
$$+ \sum_{j=1}^{2} d^{(j)}(\boldsymbol{l}+\boldsymbol{e}_j,\boldsymbol{\varphi}) b(k-1;\boldsymbol{l}+\boldsymbol{e}_j,\boldsymbol{\varphi}) \mathbb{1}_{\{k>0\}} + \sum_{\boldsymbol{\psi}\in\mathcal{S}} u(\boldsymbol{l},\boldsymbol{\psi},\boldsymbol{\varphi}) b(k;\boldsymbol{l},\boldsymbol{\psi}), \qquad (k;\boldsymbol{l},\boldsymbol{\varphi}) \in \mathbb{N}^3 \times \mathcal{S}. \quad (7)$$

The resulting set of equations now forms a recursive scheme with respect to the partial ordering $\prec$ of the vectors $(k;\boldsymbol{l},\boldsymbol{\varphi})$, where $(k;\boldsymbol{l},\boldsymbol{\varphi}) \prec (\widehat{k};\widehat{\boldsymbol{l}},\widehat{\boldsymbol{\varphi}})$ if

$$\Big[k + |\boldsymbol{l}| < \widehat{k} + |\widehat{\boldsymbol{l}}|\Big] \text{ or } \Big[k + |\boldsymbol{l}| = \widehat{k} + |\widehat{\boldsymbol{l}}| \wedge k < \widehat{k}\Big].$$

Indeed, we see that (7) expresses the coefficients $b(k;\boldsymbol{l},\boldsymbol{\varphi})$ in terms of coefficients of lower order than $(k;\boldsymbol{l},\boldsymbol{\varphi})$ with respect to $\prec$, except for the coefficient $b(k;\boldsymbol{l},\boldsymbol{\psi})$ in the last line. Therefore, the coefficients $b(k;\boldsymbol{l},\boldsymbol{\varphi})$ can be calculated recursively in increasing order with respect to $\prec$, where for each combination $(k;\boldsymbol{l})$ a set of at most $|S|$ linear equations must be solved. This set of equations generally possesses a unique solution. The only exception is when the system is totally empty ($\boldsymbol{l} = \mathbf{0}$), and thus all departure rates vanish. For $\boldsymbol{l} = \mathbf{0}, \boldsymbol{\varphi} \in \mathcal{S}$, the set of equations (7) reduces to

$$\sum_{\boldsymbol{\psi}\in\mathcal{S}} u(\mathbf{0},\boldsymbol{\varphi},\boldsymbol{\psi}) b(k;\mathbf{0},\boldsymbol{\varphi}) = \sum_{\boldsymbol{\psi}\in\mathcal{S}} u(\mathbf{0},\boldsymbol{\psi},\boldsymbol{\varphi}) b(k;\mathbf{0},\boldsymbol{\psi}) + y(k;\boldsymbol{\varphi}), \quad (8)$$

where

$$y(k;\boldsymbol{\varphi}) = -\sum_{j=1}^{2} a^{(j)}(\mathbf{0},\boldsymbol{\varphi}) b(k-1;\mathbf{0},\boldsymbol{\varphi}) \mathbb{1}_{\{k>0\}} + \sum_{j=1}^{2} d^{(j)}(\boldsymbol{e}_j,\boldsymbol{\varphi}) b(k-1;\boldsymbol{e}_j,\boldsymbol{\varphi}) \mathbb{1}_{\{k>0\}}.$$

By summing the equations of (8) over all $\boldsymbol{\varphi} \in \mathcal{S}$, we observe that these are dependent sets of equations for the coefficients $b(k;\mathbf{0},\boldsymbol{\varphi})$. The dependent sets are not contradictory, since we have that $\sum_{\boldsymbol{\varphi}\in\mathcal{S}} y(k;\boldsymbol{\varphi}) = 0$, due to a necessary balance between the empty states and the states with one product in the system. However, due to the dependence, additional equations are needed. The law of total probability provides an additional equation between the coefficients $b(k;\boldsymbol{l},\boldsymbol{\varphi})$, for $(k;\boldsymbol{l},\boldsymbol{\varphi}) \in \mathbb{N}^3 \times \mathcal{S}$ when the system is empty. Namely, observe that if we put $\chi = 0$ in (6), which corresponds to zero arrival rates, all terms vanish except for the one for $k = 0$. Thus, from the law of total probability (i.e., the normalisation equation (5)), we have

$$\sum_{\boldsymbol{\varphi}\in\mathcal{S}} b(0;\mathbf{0},\boldsymbol{\varphi}) = \sum_{\boldsymbol{\varphi}\in\mathcal{S}} p(\mathbf{0},\boldsymbol{\varphi}) = \sum_{(\boldsymbol{l},\boldsymbol{\varphi})\in\mathbb{N}^2\times\mathcal{S}} p(\boldsymbol{l},\boldsymbol{\varphi}) = 1, \quad (9)$$

where the first equality follows from (6) and the second equality follows due to the fact that if all arrival rates are zero, then the only transition probabilities are the $p(\mathbf{0},\boldsymbol{\varphi})$. Similarly, (5) implies for higher orders of $k > 0$ that

$$\sum_{\boldsymbol{\varphi}\in\mathcal{S}} b(k;\mathbf{0},\boldsymbol{\varphi}) = -\sum_{0<|\boldsymbol{l}|\le k} \sum_{\boldsymbol{\psi}\in\mathcal{S}} b(k-|\boldsymbol{l}|;\boldsymbol{l},\boldsymbol{\psi}). \quad (10)$$

In order to see how (10) is derived, we argue as follows. First, we substitute (6) into (5), and thus write the normalisation equation as a power series in χ. As this equation needs to be true for all values of χ, we have that the coefficients of χ for all powers of χ need to be equal to zero. Last, observe that (10) gives actually the coefficient for the k-th power in this series. The only modification is that the terms in the k-th coefficient that correspond to an empty system have been accumulated in the left-hand side of (10), while the remaining terms appear in the right-hand side.

Note now that the right hand side of (10) consists of terms of lower order than $b(k;\mathbf{0},\boldsymbol{\varphi})$ with respect to $\prec$. All but one of the equations of (8) in combination with (9) or (10) determine $b(k;\mathbf{0},\boldsymbol{\varphi})$. In general, this set of equations has a unique solution, if the process, conditioned on the event that both queues are empty and no arrivals occur at all, is irreducible on the subset of $\mathcal{S}$ of reachable states. This condition holds for the current model, as the Markov process $\{\boldsymbol{\Phi}(t), t \geq 0\}$ on the state space $\mathcal{S} = \{(U,U),(U,R),(R,U),(R,W),(W,R)\}$ is evidently irreducible.

With the equations above, one can now compute all the coefficients $b(k;\boldsymbol{n},\boldsymbol{\varphi})$, for $k \in \mathbb{N}$, $(\boldsymbol{n},\boldsymbol{\varphi}) \in \mathbb{N}^2 \times \mathcal{S}$ recursively. This not only allows for the computation of the steady-state probabilities, but also of any function of the state probabilities. More specifically, let $g(\boldsymbol{l},\boldsymbol{\varphi})$ represent a function which maps values from the state space $\mathbb{N}^2 \times \mathcal{S}$ to a real value. Most common performance measures, including moments of the queue lengths, can be expressed in the form $\mathbb{E}[g(\boldsymbol{L},\boldsymbol{\Phi})]$. Using (6), the expectation of $g(\boldsymbol{L},\boldsymbol{\Phi})$ is defined as

$$\mathbb{E}[g(\boldsymbol{L},\boldsymbol{\Phi})] = \sum_{(\boldsymbol{l},\boldsymbol{\varphi})\in\mathbb{N}^2\times\mathcal{S}} g(\boldsymbol{l},\boldsymbol{\varphi})p(\boldsymbol{l},\boldsymbol{\varphi}) = \sum_{m=0}^{\infty}\sum_{|\boldsymbol{l}|=m}\sum_{\boldsymbol{\varphi}\in\mathcal{S}} g(\boldsymbol{l},\boldsymbol{\varphi})\sum_{k=0}^{\infty}\chi^{k+m}b(k;\boldsymbol{l},\boldsymbol{\varphi}).$$

By changing the index of the last sum, substituting k for $k-m$, and subsequently changing the order of summation we obtain

$$\mathbb{E}[g(\boldsymbol{L},\boldsymbol{\Phi})] = \sum_{k=0}^{\infty}\chi^k\sum_{m=0}^{k}\sum_{|\boldsymbol{l}|=m}\sum_{\boldsymbol{\varphi}\in\mathcal{S}} g(\boldsymbol{l},\boldsymbol{\varphi})b(k-m;\boldsymbol{l},\boldsymbol{\varphi}).$$

This implies that performance measures of the form $\mathbb{E}[g(\boldsymbol{L},\boldsymbol{\Phi})]$ can also be written as a power series in χ:

$$\mathbb{E}[g(\boldsymbol{L},\boldsymbol{\Phi})] = \sum_{k=0}^{\infty}\chi^k f(k), \tag{11}$$

with coefficients given by

$$f(k) := \sum_{0\leq|\boldsymbol{l}|\leq k}\sum_{\boldsymbol{\varphi}\in\mathcal{S}} g(\boldsymbol{l},\boldsymbol{\varphi})b(k-|\boldsymbol{l}|;\boldsymbol{l},\boldsymbol{\varphi}), \qquad k = 0,1,\ldots \tag{12}$$

While the computation of $\mathbb{E}[g(\boldsymbol{L},\boldsymbol{\Phi})]$ involves the computation of an infinite number of coefficients, in practice only a finite number of coefficients can be computed. Since the term $\chi^k f(k)$ often converges to zero as $k \to \infty$, we can compute $\mathbb{E}[g(\boldsymbol{L},\boldsymbol{\Phi})]$ up to arbitrary precision by truncating the series after a finite number M. We thus obtain the following computational scheme to evaluate $\mathbb{E}[g(\boldsymbol{L},\boldsymbol{\Phi})]$:

1. Determine $b(0;\mathbf{0},\boldsymbol{\varphi})$ by solving the set of equations consisting of all but one of the equations (8) together with (9). Compute $f(0)$ according to (12), i.e.
$$f(0) = \sum_{\boldsymbol{\varphi}\in\mathcal{S}} g(\mathbf{0},\boldsymbol{\varphi})b(0;\mathbf{0},\boldsymbol{\varphi}). \tag{13}$$
2. Let $f(k) := 0$, $k = 1,2,\ldots$
3. Set $m := 1$.
4. For all $(k;\boldsymbol{l},\boldsymbol{\varphi}) \in \mathbb{N}^3 \times \mathcal{S}$ with $\boldsymbol{l} \neq \mathbf{0}$ and with $k + |\boldsymbol{l}| = m$, compute $b(k;\boldsymbol{l},\boldsymbol{\varphi})$ by iteratively solving the equation set (7) in increasing order of $(k;\boldsymbol{l},\boldsymbol{\varphi})$ with respect to $\prec$. Update $f(m)$ according to (12).
5. For all $\boldsymbol{\varphi} \in \mathcal{S}$, compute $b(k;\boldsymbol{l},\boldsymbol{\varphi})$ by solving the set of equations consisting of all but one of the equations (7) in combination with (10). Update $f(m)$ according to (12).
6. Set $m := m + 1$. If $m \leq M$, return to step 4, otherwise stop.

With this computational scheme, performance measures such as the r-th moment of L_i or the cross-moment $\mathbb{E}[L_1L_2]$ can be computed by taking $g(\boldsymbol{l},\boldsymbol{\varphi}) = l_i^r$ or $g(\boldsymbol{l},\boldsymbol{\varphi}) = l_1l_2$ respectively. Moreover, note that the steady-state probabilities $p(\boldsymbol{n},\boldsymbol{\psi})$ themselves can be computed through this scheme by taking $g(\boldsymbol{l},\boldsymbol{\varphi}) = \mathbb{1}_{\{\boldsymbol{l}=\boldsymbol{n},\boldsymbol{\varphi}=\boldsymbol{\psi}\}}$. We end this section with several remarks.

Remark 3.1. For the numerical evaluation of the performance measures, we compute (11) using the corresponding function $g(\boldsymbol{l}, \boldsymbol{\varphi})$ and truncate the power series after the M-th order term. In general, it is hard to say exactly how to choose the value of M in order to achieve a certain degree of accuracy. First, this number depends on the 'degree of symmetry'. If the rates of arrival, service, breakdown and repair do not differ between the first-layer queues and machines, the power series (11) generally converges faster than for systems, where these rates are queue-dependent or machine-dependent. Secondly, the choice of M also depends heavily on the load offered to the system. For small χ, only a small number of terms have to be computed for the truncated power series to be accurate.

Remark 3.2. It is not guaranteed that the power series (6) and (11) converge for every value of χ. Therefore, it may happen that PSA, as presented in this section, will fail for very asymmetric systems, because (6) and (11) are divergent. There are two techniques available in the literature to improve the convergence properties of these power series. For an extensive discussion of these methods, see e.g. [5]. The conformal mapping technique tries to enlarge the radius of convergence by mapping any singularities out of the circle $|\chi| < \chi^*$. Alternatively, the epsilon algorithm accelerates convergence of a slowly convergent power series or determines a value for a divergent series. This is done by approximating the performance measure under consideration by a sequence of polynomials.

Remark 3.3. Observe that we have assumed exponentiality in the interarrival times, service times, breakdown times and repair times. However, this is not strictly needed to apply PSA. In order to use PSA, we only need phase-type distributions. For phase-type distributions, the supplementary vector $\boldsymbol{\Phi}(t)$ must be expanded to include information on the phase each of the running times is in, in order to preserve the Markov property of the process $\{(\boldsymbol{L}(t), \boldsymbol{\Phi}(t)), t \geq 0\}$. Therefore, the size of the supplementary state space $\mathcal{S}$ increases. This may lead to a considerable increase in complexity of the computational scheme, since the equation set (7) now contains more equations and more unknowns. For Coxian distributions however, the increased complexity is limited, since the phases of a Coxian distribution are placed in sequence. Therefore, (7) will be a relatively sparse set of equations.

Remark 3.4. In this section, we have applied PSA to a layered model with two machines (resulting in two first-layer queues) and one repairman. PSA is also applicable to a similar model with a larger number of machines and repairmen. For a larger number of machines and first-layer queues, information on the order in which the machines are waiting for repair needs to be included into the supplementary vector $\boldsymbol{\Phi}(t)$. Because the dimension of the vector $\boldsymbol{L}(t)$ and the size of the state space $\mathcal{S}$ will increase, the computational complexity will increase accordingly. For a larger number of repairmen, no additional non-exponentiality is introduced to the system and thus no additional information needs to be included into $\boldsymbol{\Phi}(t)$, however the state space $\mathcal{S}$ and the rates $u(\boldsymbol{l}, \boldsymbol{\varphi}, \boldsymbol{\psi})$ will evidently change.

4 Light traffic behaviour

In Section 3, we have derived a computational scheme to numerically compute performance measures. These computations can be performed up to arbitrary precision, by truncating the power series in (11) and subsequently computing recursively the coefficients $f(k)$. This leads to the question whether PSA can also be used to obtain similar computations in a *symbolic* fashion. In theory, this is possible by running the computational scheme as before, but now using parameter values instead of numerical values for the rates of arrival, service, breakdown and repair. However, from a practical point of view, only coefficients $f(k)$ up to a small number of k can be computed symbolically. The set of equations (7) becomes increasingly hard to solve, as the expressions for the terms $b(k; \boldsymbol{l}, \boldsymbol{\varphi})$ become very large very fast.

The number of coefficients that can be computed symbolically is generally not enough to obtain an accurate approximation for general values of χ. However, as χ becomes smaller, the higher-order terms become increasingly negligible. Therefore, the so-called light-traffic (LT) behaviour of a performance measure as $\chi \to 0$ can be identified symbolically. We do so for the performance measures $\mathbb{E}[L_1]$ in Section 4.1 and $\mathbb{E}[L_1 L_2]$ in Section 4.2. For the sake of clarity, we will refer to the k-th coefficients $f(k)$ in (11) corresponding to $g(\boldsymbol{l}, \boldsymbol{\varphi}) = l_1$ as $f_1(k)$ in the sequel. Similarly, $f_2(k)$ denotes the k-th coefficient corresponding to $g(\boldsymbol{l}, \boldsymbol{\varphi}) = l_1 l_2$.

4.1 Marginal queue length

We are interested in the LT behaviour of the marginal queue length L_1 in the variable $\chi = \hat{\rho}$. More specifically, we regard the behaviour of the mean of L_1 as a function of the relative load, as $\hat{\rho}$ goes to zero. By taking $g(\boldsymbol{l}, \boldsymbol{\varphi}) = l_1$, and running PSA with $M = 2$, we obtain the following expression for $\mathbb{E}[g(\boldsymbol{L}, \boldsymbol{\Phi})] = \mathbb{E}[L_1]$:

$$\mathbb{E}[L_1] = f_1(0) + f_1(1)\hat{\rho} + f_1(2)\hat{\rho}^2 + \mathcal{O}(\hat{\rho}^3), \tag{14}$$

where $\mathcal{O}(\hat{\rho}^3)$ represents third- and higher order terms in $\hat{\rho}$ for which the computation of the coefficients is not feasible symbolically. Furthermore, we have that $f_1(0) = 0$, since $g(\mathbf{0}, \boldsymbol{\varphi}) = 0^r = 0$ in (13). This is explained by the fact that there are no arrivals for $\hat{\rho} = 0$, and thus there never is any product in the system. The coefficient $f_1(1)$ equals $\frac{d}{d\hat{\rho}}\mathbb{E}[L_1]|_{\hat{\rho}\downarrow 0}$, the derivative of the mean of L_1 with respect to $\hat{\rho}$ evaluated at $\hat{\rho} \downarrow 0$. Computing $f_1(1)$ leads to a closed-form expression in the service rate of M_1 as well as the breakdown and repair rates of each of the machines. Since this term is too large to display in full, we give the expressions for $\frac{d}{d\hat{\rho}}\mathbb{E}[L_1]|_{\hat{\rho}\downarrow 0}$ in each of the model parameters separately. When giving the derivative in each rate of these parameters, we assume all other parameters to be equal to one:

Model parameter	$\frac{d}{d\hat{\rho}}\mathbb{E}[L_1]\|_{\hat{\rho}\downarrow 0} = f_1(1)$
μ_1	$\frac{26}{25} + \frac{8\mu_1}{25} - \frac{3}{25(3+\mu_1)}$
σ_1	$1 + \frac{9}{49(3+\sigma_1)} - \frac{36}{7(2+3\sigma_1)^2} + \frac{120}{49(2+3\sigma_1)}$
σ_2	$\frac{4}{3} - \frac{3}{49(3+\sigma_2)} - \frac{13}{21(2+3\sigma_2)^2} + \frac{9}{49(2+3\sigma_2)}$
ν_1	$\frac{75}{64} + \frac{135}{256(1+2\nu_1)^2} + \frac{21}{256(1+2\nu_1)} + \frac{567}{256(3+2\nu_1)^2} - \frac{21}{256(3+2\nu_1)}$
ν_2	$\frac{5}{4} - \frac{13}{75(3+\nu_2)} + \frac{27}{20(1+2\nu_2)^2} - \frac{57}{100(1+2\nu_2)} - \frac{1}{2(3+2\nu_2)^2} + \frac{11}{12(3+2\nu_2)}$

From these results, we see that $\frac{d}{d\hat{\rho}}\mathbb{E}[L_1]|_{\hat{\rho}\downarrow 0}$ is increasing in μ_1, and decreasing in ν_1 and ν_2. This is not surprising, as it intuitively makes sense that the queue length generally increases in the service rates and decreases in the repair rates. Moreover, we note that the denominators of the terms in the expressions only involve the model parameters in the form of polynomials of at most second order; there are no higher powers involved.

It is important to observe that the expression $f_1(1)$ also represents the first-order derivative of higher moments of L_1 as $\hat{\rho} \downarrow 0$. In other words, the first-order derivative of $\mathbb{E}[L_i^r]$ with respect to $\hat{\rho}$, evaluated at $\hat{\rho} \downarrow 0$ is independent of r. This can be explained by careful inspection of $f(1)$ in (12). The first-order term $f_1(1)$ only involves values of $g(\boldsymbol{l}, \boldsymbol{\varphi})$ for which $|\boldsymbol{l}| \in \{0, 1\}$, which implies that $l_1 \in \{0, 1\}$ too. To inspect $\mathbb{E}[L_i^r]$, we take $g(\boldsymbol{l}, \boldsymbol{\varphi}) = l_1^r$. Since $l_1 \in \{0, 1\}$, the function $g(\boldsymbol{l}, \boldsymbol{\varphi})$ can only evaluate to the values $0^r = 0$ or $1^r = 1$, irrespective of $r > 0$.

The application of PSA in a symbolic manner also allows us to find closed-form expressions for the second-order derivative $\frac{d^2}{d\hat{\rho}^2}\mathbb{E}[L_1]|_{\hat{\rho}\downarrow 0} = 2f_1(2)$. Again, we give this expression in each of the model parameters, while assuming the other parameters to be equal to one:

Model parameter	$\frac{d^2}{d\hat{\rho}^2}\mathbb{E}[L_1]\|_{\hat{\rho}\downarrow 0} = 2f_1(2)$
μ_1	$\frac{226}{125} + \frac{88\mu_1}{125} - \frac{108}{125(3+\mu_1)^3} - \frac{18}{125(3+\mu_1)^2} + \frac{18}{25(3+\mu_1)}$
σ_1	$2 - \frac{36}{343(3+\sigma_1)^3} + \frac{222}{2401(3+\sigma_1)^2} + \frac{4538}{16807(3+\sigma_1)} + \frac{272}{343(2+3\sigma_1)^3} - \frac{26800}{2401(2+3\sigma_1)^2} + \frac{87228}{16807(2+3\sigma_1)}$
σ_2	$\frac{8}{3} - \frac{12}{343(3+\sigma_2)^3} - \frac{500}{2401(3+\sigma_2)^2} - \frac{3784}{16807(3+\sigma_2)} - \frac{68}{343(2+3\sigma_2)^3} - \frac{10952}{7203(2+3\sigma_2)^2} + \frac{11352}{16807(2+3\sigma_2)}$
ν_1	$\frac{2385}{1024} - \frac{459}{8192(1+2\nu_1)^3} + \frac{22725}{16384(1+2\nu_1)^2} - \frac{11673}{16384(1+2\nu_1)} + \frac{19683}{8192(3+2\nu_1)^3} + \frac{105309}{16384(3+2\nu_1)^2} + \frac{12249}{16384(3+2\nu_1)}$
ν_2	$\frac{5}{2} - \frac{52}{1125(3+\nu_2)^3} - \frac{2312}{5625(3+\nu_2)^2} - \frac{48194}{84375(3+\nu_2)} - \frac{1107}{4000(1+2\nu_2)^3} + \frac{110367}{40000(1+2\nu_2)^2} - \frac{106017}{100000(1+2\nu_2)} - \frac{283}{288(3+2\nu_2)^3} - \frac{59}{64(3+2\nu_2)^2} + \frac{1903}{864(3+2\nu_2)}$

Again, we see that $\frac{d^2}{d\hat{\rho}^2}\mathbb{E}[L_1]|_{\hat{\rho}\downarrow 0}$ is increasing in μ_1, and decreasing in ν_1 and ν_2. Furthermore, note that the denominators of the expressions only involve the model parameters in a polynomial fashion up to order three.

This is not surprising, as the expressions for the first derivative only involve the parameters up to a second order.

Remark 4.1. If we wish to compute the LT behaviour of the moments of L_2, we perform similar computations to the above, or we simply renumber the queues.

Remark 4.2. Note that the computation of $f_1(1) = \frac{d}{d\hat{\rho}}\mathbb{E}[L_1]|_{\hat{\rho}\downarrow 0}$ is also possible through Little's law:

$$\mathbb{E}[S_1]|_{\hat{\rho}\downarrow 0} = \frac{\mathbb{E}[L_1]|_{\hat{\rho}\downarrow 0}}{\lambda_1} = \frac{f_1(1)\hat{\rho} + \mathcal{O}(\hat{\rho}^2)}{\lambda_1}\Big|_{\hat{\rho}\downarrow 0} = \frac{f_1(1)}{\mu_1 m_C} = \frac{\frac{d}{d\hat{\rho}}\mathbb{E}[L_1]|_{\hat{\rho}\downarrow 0}}{\mu_1 m_C}, \tag{15}$$

where $\mathbb{E}[S_1]|_{\hat{\rho}\downarrow 0}$ is the mean sojourn time of a type-1 product conditioned on the event there are no other products in the system. This sojourn time consists of the actual service requirement, the time the product needs to wait before M_1 takes the product into service, and the downtime M_1 suffers during the service of the product. The mean of the first term obviously equals μ_1^{-1}. The means of the latter two terms can be computed by studying the Markov process $\{\mathbf{\Phi}(t), t > 0\}$. Eventually, this leads to an expression for $\mathbb{E}[S_1]|_{\hat{\rho}\downarrow 0}$, which on its turn leads to an expression for $\frac{d}{d\hat{\rho}}\mathbb{E}[L_1]|_{\hat{\rho}\downarrow 0}$ due to (15).

4.2 Joint queue length

In this section, we discuss the LT behaviour of $\mathbb{E}[L_1 L_2]$, the cross-moment of the queue lengths in the layered model, as a function of $\hat{\rho}$. We regard instances of the model for which both of the arrival rates tend to zero, while we preserve the relative values, i.e., we assume that

$$\lambda_2 = d\lambda_1$$

at all times for a constant $d > 0$. This means that we set $a^{(1)}(\boldsymbol{l}, \boldsymbol{\varphi}) = \mu_1 m_C$ and $a^{(2)}(\boldsymbol{l}, \boldsymbol{\varphi}) = d\mu_1 m_C$, while we let λ_1 (or $\hat{\rho}$) go to zero. Furthermore, we take $g(\boldsymbol{l}, \boldsymbol{\varphi}) = l_1 l_2$. By running the computational scheme as given in Section 3.2 with $M = 2$, we obtain the following expression for $\mathbb{E}[L_1 L_2]$:

$$\mathbb{E}[L_1 L_2] = f_2(0) + f_2(1)\hat{\rho} + f_2(2)\hat{\rho}^2 + \mathcal{O}(\hat{\rho}^3). \tag{16}$$

Like before, we have that $f_2(0) = 0$, because $g(\mathbf{0}, \boldsymbol{\varphi}) = 0$ for all $\varphi \in \mathcal{S}$ in (13). We also have that $f_2(1) = 0$, due to (12). The coefficient $f_2(1)$ only involves values of $g(\boldsymbol{l}, \boldsymbol{\varphi})$ for which $0 \leq l_1 + l_2 \leq 1$. Within this domain, there is no combination (l_1, l_2) for which $l_1 l_2 > 0$. Therefore, the most prominent LT behaviour is captured by the term $f_2(2)$.

Going back to the derivatives of the cross-moment, we have that the first order derivative of $\mathbb{E}[L_1 L_2]$ for $\hat{\rho} \downarrow 0$ vanishes, since $f_2(1) = 0$. By (16), we have for the second-order derivative that $\frac{d^2}{d\hat{\rho}^2}\mathbb{E}[L_1 L_2]|_{\hat{\rho}\downarrow 0} = 2f_2(2)$. By evaluation of the computational scheme up to $M = 2$, we obtain a closed-form expression for this second-order derivative evaluated at $\chi \downarrow 0$. Again, we give the expression separately in each of the model parameters, while assuming each of the others to be equal to one:

Model parameter	$\frac{d^2}{d\hat{\rho}^2}\mathbb{E}[L_1 L_2]\|_{\hat{\rho}\downarrow 0} = 2f_2(2)$
μ_1	$-\frac{1816d}{3375} + \frac{3646d\mu_1}{1125} + \frac{413d\mu_1^2}{375} + \frac{36d}{125(3+\mu_1)} + \frac{32(373d+143d\mu_1)}{3375\left(8+13\mu_1+3\mu_1^2\right)}$
μ_2	$\frac{413d}{375} + \frac{133d}{50\mu_2} + \frac{4d}{125(3+\mu_2)} + \frac{4357d+1235d\mu_2}{750\left(8+13\mu_2+3\mu_2^2\right)}$
σ_1	$-\frac{20d}{1+\sigma_1} + \frac{96d}{343(3+\sigma_1)^2} - \frac{3560d}{2401(3+\sigma_1)} - \frac{120d}{2197(5+\sigma_1)} - \frac{110360d}{1911(2+3\sigma_1)^3}$ $-\frac{5295776d}{173901(2+3\sigma_1)^2} + \frac{425228092d}{5274997(2+3\sigma_1)}$
σ_2	$\frac{97d}{27} + \frac{4d\sigma_2}{3} + \frac{96d}{343(3+\sigma_2)^2} - \frac{4232d}{2401(3+\sigma_2)} - \frac{480d}{2197(5+\sigma_2)}$ $-\frac{27590d}{17199(2+3\sigma_2)^3} - \frac{235012d}{57967(2+3\sigma_2)^2} - \frac{5574427d}{5274997(2+3\sigma_2)}$
ν_1	$\frac{4779d}{896} - \frac{364d}{125(3+\nu_1)} + \frac{3267d}{5120(1+2\nu_1)^3} + \frac{357131d}{51200(1+2\nu_1)^2} - \frac{3557887d}{6912000(1+2\nu_1)}$ $-\frac{189855d}{13312(3+2\nu_1)^3} + \frac{5779335d}{346112(3+2\nu_1)^2} - \frac{78477979d}{4499456(3+2\nu_1)} + \frac{17756000d}{415233(17+7\nu_1)}$
ν_2	$\frac{531d}{224} + \frac{34d}{1+\nu_2} - \frac{364d}{1125(3+\nu_2)} + \frac{3267d}{1280(1+2\nu_2)^3} + \frac{313571d}{12800(1+2\nu_2)^2} - \frac{60000667d}{1728000(1+2\nu_2)}$ $-\frac{21095d}{3328(3+2\nu_2)^3} - \frac{4655195d}{259584(3+2\nu_2)^2} - \frac{291320479d}{10123776(3+2\nu_2)} + \frac{710240d}{415233(17+7\nu_2)}$

As in the previous case, we note that the numerators and the denominators of the terms in $\frac{d^2}{d\hat{\rho}^2}\mathbb{E}[L_1L_2]|_{\hat{\rho}\downarrow 0}$ only involve the model parameters in a polynomial fashion up to order two. For the service rates μ_1 and μ_2, the expressions are equivalent. If we let λ_2 scale along with λ_1 such that $\rho_1 = \rho_2$, we even have that the expressions for μ_1 and μ_2 in the table above are the same. In this case, it holds that $\hat{\rho} = \frac{\lambda_1}{\mu_1 m_C} = \frac{\lambda_2}{\mu_2 m_C}$ and $d = \mu_2/\mu_1$. The parameter $\hat{\rho}$, thus, depends on the service rates μ_1 and μ_2 in the same way. Hence, by assuming $d = \mu_2/\mu_1$, we have that $\frac{d^2}{d\hat{\rho}^2}\mathbb{E}[L_1^T]|_{\hat{\rho}\downarrow 0}$ also depends on μ_1 and μ_2 in a similar manner. Also the corresponding equations for the breakdown rates σ_1 and σ_2, as well as those for the repair rates ν_1 and ν_2, are equivalent. When multiplying these expressions with m_C^{-2}, the results for σ_1 and σ_2 are exactly the same, as well as those for ν_1 are ν_2. This is not surprising, because $\mathbb{E}[L_1L_2]$ behaves symmetrically with respect to both of the queue lengths, and is therefore equally sensitive to characteristics of either of the machines.

5 Approximations for the mean queue length

Based on the LT-behaviour found in Section 4, we propose two approximations for $\mathbb{E}[L_1]$, the mean queue length of Q_1. Note that this does not imply a loss of generality, as one can simply renumber the queues in order to study Q_2. The first one is based on $f_1(0)$, $f_1(1)$ and $f_1(2)$, i.e. the first few coefficients of $f(k)$ in (11) corresponding to $g(\boldsymbol{l}, \boldsymbol{\varphi}) = l_1$ up to $k = 2$. This approximation is derived in Section 5.1.1. A numerical study in Section 5.1.2 shows that the approximation already achieves a very good accuracy. In the worst case of the model instances tested, the relative difference between the approximated mean queue length and the simulated value is 1.7%. Moreover, since the coefficients $f_1(0)$, $f_1(1)$ and $f_1(2)$ are still tractable in a symbolic fashion, as we have seen in Section 4.1, the approximation is in closed form. Therefore, it is easily implementable and suitable for optimisation purposes. The first approximation is only based on LT-results obtained by the application of PSA.

In an effort to further increase accuracy, we also study the heavy-traffic (HT) behaviour in Section 5.2.1; i.e., the behaviour of the mean queue length as the queue becomes fully saturated. We subsequently propose a second approximation for the mean queue length in Section 5.2.2, based on an interpolation between the same LT-limits already used in the closed-form approximation and the HT behaviour of the queue length. This second approximation is therefore also exact when the queue is fully saturated. The interpolation approximation works even better in terms of accuracy, being indistinguishable from simulation results. However, it is not in closed form, and is thus slightly harder to implement. Finally, we present a number of limiting cases in Section 5.3, where both approximations turn out to be exact.

5.1 Closed-form approximation

In this section, we derive a closed-form approximation for the mean queue length of Q_1 based on symbolic closed-form expressions of the first three PSA-coefficients $f_1(0)$, $f_1(1)$ as well as $f_1(2)$, which we denote by $\mathbb{E}[L_{1,app}^{\text{CF}}]$ and then numerically assess its accuracy.

5.1.1 Derivation of the closed-form approximation

We first develop a closed-form approximation for $\mathbb{E}[L_1]$, by observing the mean queue length as a function h of $\hat{\rho}$. We assume this function to be analytic on $[0, 1)$; i.e., we have that

$$h(\hat{\rho}) = \sum_{n=0}^{\infty} f_1(n)\hat{\rho}^n = \frac{z(\hat{\rho})}{1-\hat{\rho}} \qquad 0 \leq \hat{\rho} < 1, \tag{17}$$

where

$$f_1(n) = \frac{h^{(n)}(0)}{n!}, \qquad z(\hat{\rho}) = f_1(0) + \sum_{n=1}^{\infty}(f_1(n) - f_1(n-1))\hat{\rho}^n \tag{18}$$

Parameter	Considered parameter values
ρ_1	$\{0.25, 0.5, 0.75\}$
δ_1	$\{1\}$
(σ_1, σ_2)	$a_i^\sigma \cdot b_j^\sigma \qquad \forall i,j$ where $\boldsymbol{a}^\sigma = \{0.1, 1, 10\}$ and $\boldsymbol{b}^\sigma = \{(1,1),(1,2),(2,1),(1,5),(5,1)\}$
(ν_1, ν_2)	$a_i^\nu \cdot b_j^\nu \qquad \forall i,j$ where $\boldsymbol{a}^\nu = \{0.1, 1, 10\}$ and $\boldsymbol{b}^\nu = \{(1,1),(1,2),(2,1),(1,5),(5,1)\}$

Table 1: Parameter values of the test bed used to compare the closed-form approximation to exact results.

and $h^{(n)}(\hat{\rho})$ is the n-th derivative of h with respect to $\hat{\rho}$. Note that the power series (11) and (17) are equal when taking $g(\boldsymbol{l}, \boldsymbol{\varphi}) = l_1$ and $\chi = \hat{\rho}$. As a consequence, although an exact expression for $h(\hat{\rho})$ is not known, the coefficients $f_1(n)$, $n = 0, 1, \ldots$ can be computed using the computational scheme as given in Section 3.2. Numerically, we can do this for large n. Symbolically, however, only $f_1(0) = 0, f_1(1)$ and $f_1(2)$ lead to tractable closed-form expressions in the model parameters $\mu_1, \sigma_1, \sigma_2, \nu_1$ and ν_2. Note that we have already studied these expressions in detail in Section 4.1. Since $h(\hat{\rho})$ is guaranteed to exist, we have that the sum in $z(\hat{\rho})$ must converge. When observing numerically the first few terms of the series $\{f_1(i) - f_1(i-1), i > 0\}$ using the computational scheme of Section 3.2, we generally see that they are moderate in absolute value, but more importantly, alternate in sign rapidly. This even seems to be the case when this series is divergent. Because of this, and the decreasing nature of $\hat{\rho}^n$ in n, we may assume that the first two terms alone already approximate this sum well. In other words, since $f_1(0)$ is 0, we have that $z(\hat{\rho})$ is well approximated by $f_1(1)\hat{\rho} + (f_1(2) - f_1(1))\hat{\rho}^2$. Out of this observation, a closed-form approximation for $\mathbb{E}[L_1]$ follows immediately.

Approximation 5.1. *In the two-layered model, an accurate closed-form approximation for the mean queue length of Q_i is given by*

$$\mathbb{E}[L_{1,app}^{\text{CF}}] = \frac{a\hat{\rho} + b\hat{\rho}^2}{1 - \hat{\rho}}, \tag{19}$$

where $a = f_1(1)$, $b = f_1(2) - f_1(1)$ and $\hat{\rho} = \frac{\lambda_1}{\mu_1(\boldsymbol{\pi}^{\boldsymbol{\Phi}}_{(U,U)} + \boldsymbol{\pi}^{\boldsymbol{\Phi}}_{(U,R)})}$.

An extensive numerical study in the next section shows that Approximation 5.1 performs very well in terms of accuracy. Furthermore, because the approximation is given in a simple and closed form, it is very easy to implement and suitable for optimisation purposes.

5.1.2 Accuracy of the closed-form approximation

In this section, we apply Approximation 5.1 on a number of systems and compare it to "exact" values of the mean queue length of Q_1, obtained by simulation. The complete test bed of instances that we analysed contains 675 different combinations of parameter values, all listed in Table 1. This table lists multiple values for the normalised workload of $Q_1(\hat{\rho})$, the breakdown rates of M_1 and M_2 (σ_1 and σ_2) and the repair rates of M_1 and M_2 (ν_1 and ν_2). In particular, these rates are varied in the order of magnitude through the values a_i^σ and a_i^ν and in the imbalance, through the values b_j^σ and b_j^ν, all specified in the table. As a consequence, the breakdown rates (σ_1, σ_2) and the repair rates (ν_1, ν_2) run from $(0.1, 0.1)$, being small and perfectly balanced, to $(50, 10)$, being large and significantly imbalanced. The service requirements of type-1 products are assumed to be exponentially (1) distributed.

For each of the systems corresponding to each of the parameter combinations in Table 1, we compare the *approximated* mean queue length of Q_1, $\mathbb{E}[L_{1,app}^{\text{CF}}]$, to the "exact" mean queue length $\mathbb{E}[L_1]$. Subsequently, we compute

the relative error of these approximations, i.e.,

$$\Delta := 100\% \times \Big| \frac{\mathbb{E}[L_{1,app}^{\text{CF}}] - \mathbb{E}[L_1]}{\mathbb{E}[L_1]} \Big|.$$

The average value for Δ in this test bed is roughly 0.07%. The system for which the error Δ is largest with a value of 1.7%, is given for the model parameters $\hat{\rho} = 0.75$ and $\sigma_1 = \sigma_2 = \nu_1 = \nu_2 = 0.1$, i.e., the system for which the breakdowns and repairs occur on the slowest time scale compared to the interarrival times and service times of the products in the first queue.

In Table 2, the mean value of Δ is given for each category of the variables in Table 1. We see in Table 2(a) that the accuracy of the approximation increases as the load offered to Q_1 decreases. This is not surprising, as the approximation is exact in LT by construction. From Tables 2(b) and 2(c), it is clear that the approximation is sensitive to the magnitude of the breakdown rates and repair rates. As will be evident in Section 5.3, the approximation becomes exact as some of these variables tend to zero or infinity. Moreover, according to Tables 2(d) and 2(e) the approximation is not very sensitive to imbalance in the second layer of the system.

(a)

$\hat{\rho}$	0.25	0.5	0.75
Mean rel. error	0.017%	0.064%	0.127%

(b)

a_i^σ	0.1	1	10
Mean rel. error	0.200%	0.007%	0.001%

(c)

a_i^ν	0.1	1	10
Mean rel. error	0.202%	0.005%	0.001%

(d)

b_j^σ	(1, 1)	(1, 2)	(2,1)	(1, 5)	(5, 1)
Mean rel. error	0.083%	0.073%	0.072%	0.079%	0.038%

(e)

b_j^ν	(1, 1)	(1, 2)	(2,1)	(1, 5)	(5, 1)
Mean rel. error	0.091%	0.080%	0.029%	0.045%	0.101%

Table 2: Mean relative error categorised in ρ_1 (a), a_i^σ (b), a_i^ν (c), b_j^σ (d) and b_j^ν (e).

From these results, we conclude that the approximation works very well in general. The accuracy may degrade slightly when breakdown rates and repair rates are very small compared to the arrival and service rate of type-1 products. To illustrate this, regard a system with $\mu_1 = 1$ and $\sigma_1 = \sigma_2 = \nu_1 = \nu_2 = 0.001$. In Figure 2, we plot the closed-form approximation $\mathbb{E}[L_{1,app}^{\text{CF}}]$ along with the numerical values for $\mathbb{E}[L_1]$ versus $\hat{\rho}$. In this extreme example, Δ grows up to roughly 6% as $\hat{\rho}$ nears one. However, the closed-form approximation remains very well suited for optimisation purposes. The shapes of the curves of $\mathbb{E}[L_{1,app}^{\text{CF}}]$ and $\mathbb{E}[L_1]$ still match each other well. Therefore, using the derived closed-form approximation into an optimisation function, instead of an exact expression if it had been available, should result in an optimum that is close to the true optimum.

5.2 Interpolation approximation

Approximation 5.1 satisfies the light-traffic limits found by PSA and already performs very well. To further increase performance, we refine the approximation so that it also satisfies the heavy-traffic behaviour of the mean

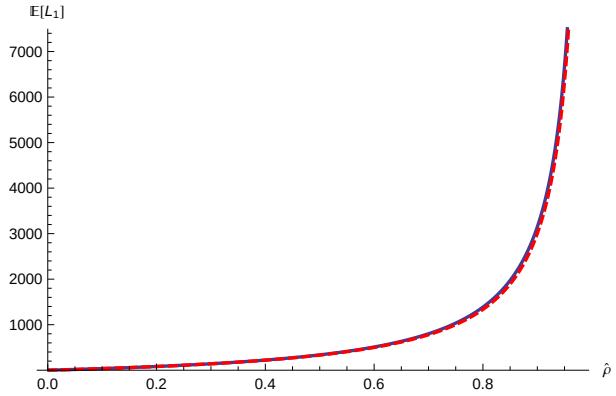

Figure 2: $\mathbb{E}[L^{\text{CF}}_{1,app}]$ (solid curve) and $\mathbb{E}[L_1]$ (dashed curve) versus $\hat{\rho}$.

queue length. More specifically, based on the form of the performance measure $h(\hat{\rho})$ in (17), we construct an approximation $\tilde{h}(\hat{\rho})$ so that it matches every limit that is known about $h(\hat{\rho})$; i.e., for light traffic

$$h^{(i)}(0) = \tilde{h}^{(i)}(0), i = 0, 1, \ldots, k-1, \tag{20}$$

and for heavy traffic

$$\lim_{\hat{\rho}\uparrow 1}(1-\hat{\rho})h(\hat{\rho}) = \lim_{\hat{\rho}\uparrow 1}(1-\hat{\rho})\tilde{h}(\hat{\rho}). \tag{21}$$

In literature [9, 18, 22], such interpolation approximations of the form

$$\tilde{h}(\hat{\rho}) = \frac{\sum_{n=0}^{k} r(n)\hat{\rho}^n}{1-\hat{\rho}} \tag{22}$$

have been proposed and used successfully to approximate performance measures in the GI/G/1 queue and in queueing systems with Poisson input. More recently, an interpolation approximation of this sort has been successfully applied to approximate the mean waiting time in polling systems with renewal arrivals [6], which has acted as a basis for a distributional waiting-time approximation in such systems [7].

Note that Approximation 5.1 already is of the form of (22), with $k = 3$, $r(0) = r(3) = 0, r(1) = f_1(1)$ and $r(2) = f_1(2) - f_1(1)$. The closed-form approximation already satisfies the LT limits; however it does not satisfy the HT limit. Therefore, we refine the approximation by taking $r(3)$ such that the approximation is also exact in heavy traffic. To this end, we first determine heavy-traffic limits in Section 5.2.1, after which we discuss the resulting interpolation approximation in Section 5.2.2.

5.2.1 Heavy-traffic limit

In this section, we present the heavy-traffic limit for the mean queue length of Q_1; i.e., $\lim_{\hat{\rho}\uparrow 1}\mathbb{E}[(1-\hat{\rho})L_1]$. This result is one of the ingredients for the intended interpolation approximation, see (21). We also provide the outline of a proof leading to this result. A complete proof is beyond the scope of the current paper and will be presented in a forthcoming paper.

Theorem 5.1. *The random variable* $\lim_{\hat{\rho}\uparrow 1}(1-\hat{\rho})L_1$ *is exponentially distributed with mean* $1 + \frac{\mu_1\sigma_C^2}{2m_C}$*, where*

$$m_C = \lim_{t\to\infty}\frac{\mathbb{E}[C(t)]}{t} = \boldsymbol{\pi}^{\boldsymbol{\Phi}}_{(U,U)} + \boldsymbol{\pi}^{\boldsymbol{\Phi}}_{(U,R)} \qquad \textit{and} \qquad \sigma_C^2 = \lim_{t\to\infty}\frac{\text{Var}[C(t)]}{t}.$$

Outline of proof. Before considering the queue length, first regard the stationary amount of work W_1 in Q_1. Let $A(\lambda_1 t)$ be the cumulative amount of work which has entered the queue in $[0, t)$, i.e.,

$$A(\lambda_1 t) = \sum_{i=1}^{N(\lambda_1 t)} B_i,$$

where $N(t)$ is a Poisson distributed random variable with rate t and the B_i are i.i.d. exponentially distributed with rate μ_1. By inspection of the one-sided reflection of the net-input process $\{A(\lambda_1 t) - C(t), t \geq 0\}$, the distribution of the stationary amount of work in Q_1 is determined by

$$W_1 \stackrel{d}{=} \sup_{t \geq 0} \{A(\lambda_1 t) - C(t)\}. \tag{23}$$

Let $R := \frac{1}{1-\hat{\rho}}$. Dividing both sides of (23) by R and scaling time by R^2, we obtain

$$\begin{aligned}(1-\hat{\rho})W_1 &\stackrel{d}{=} \sup_{t \geq 0} \left\{ \frac{A(\lambda_1 R^2 t) - C(R^2 t)}{R} \right\} \\ &= \sup_{t \geq 0} \left\{ \frac{A(\lambda_1 R^2 t) - \mathbb{E}[A(\lambda_1 R^2 t)]}{R} - \frac{C(R^2 t) - \mathbb{E}[C(R^2 t)]}{R} - \frac{\mathbb{E}[C(R^2 t)] - \mathbb{E}[A(\lambda_1 R^2 t)]}{R} \right\}.\end{aligned}$$

Taking the limit $R \to \infty$ (or equivalently, $\hat{\rho} \uparrow 1$) on both sides, one can show that we are allowed to interchange limit and supremum operators in the right hand side. Due to the functional central limit theorem [23], we subsequently have that the first two terms converge to Brownian motions with zero drift and variances $\mu_1 m_C \sigma_A^2 = \frac{2m_C}{\mu_1}$ and σ_C^2 respectively. The third term converges to $m_C t$ as $R \to \infty$. Thus, we have for the stationary amount of work in the system that the random variable $\lim_{\hat{\rho}\uparrow 1}(1-\hat{\rho})W_1$ is in distribution equal to the all-time supremum of a Brownian motion with drift $-m_C$ and variance $\frac{2m_C}{\mu_1} + \sigma_C^2$. By standard theory on the Brownian motion, this is known to be exponentially distributed with mean $\frac{1}{\mu_1} + \frac{\sigma_C^2}{2m_C}$.

Now that a HT limit for the stationary amount of work in Q_1 is known, HT limits for the waiting time D_1, the sojourn time S_1 and ultimately the queue length L_1 follow. By the HT limit for W_1 and the relation

$$\mathbb{P}(D_1 > t) = \mathbb{P}(W_1 > C(t)),$$

one can prove that $\lim_{\hat{\rho}\uparrow 1}(1-\hat{\rho})D_1$ is exponentially distributed with mean $\frac{1}{\mu_1 m_C} + \frac{\sigma_C^2}{2\mu_C^2}$. The sojourn time S_1 is composed of the waiting time D_1 and the service period E_1 (including service interruptions). The duration of the latter is independent to the load offered to the system. Therefore, the dynamics of E_1 are negligible in HT; i.e. $\lim_{\hat{\rho}\uparrow 1}(1-\hat{\rho})E_1 = 0$. Thus, $\lim_{\hat{\rho}\uparrow 1}(1-\hat{\rho})S_1$ follows the same distribution as $\lim_{\hat{\rho}\uparrow 1}(1-\hat{\rho})D_1$. Finally, an application of the distributional form of Little's law (cf. [15]) leads to the theorem. □

5.2.2 Resulting interpolation approximation

Now that the HT limit of the mean queue length is known, we finalise the construction of the interpolation approximation. In order to satisfy the known limiting regimes, we impose several constraints on the interpolation approximation. First, as stated in (20), we require the approximated mean waiting time at $\hat{\rho} = 0$, as well as its first two derivatives with respect to $\hat{\rho}$ evaluated at that point, to be equal to the corresponding exact values obtained by PSA:

1. $\mathbb{E}[L_{1,app}^{\text{IP}}]|_{\hat{\rho}=0} = \mathbb{E}[L_1]|_{\hat{\rho}=0} = f_1(0) = 0$,
2. $\frac{d}{d\hat{\rho}}\mathbb{E}[L_{1,app}^{\text{IP}}]|_{\hat{\rho}=0} = \frac{d}{d\hat{\rho}}\mathbb{E}[L_1]|_{\hat{\rho}=0} = f_1(1)$,
3. $\frac{d^2}{d\hat{\rho}^2}\mathbb{E}[L_{1,app}^{\text{IP}}]|_{\hat{\rho}=0} = \frac{d^2}{d\hat{\rho}^2}\mathbb{E}[L_1]|_{\hat{\rho}=0} = 2f_1(1) + 2f_1(2)$.

The terms $f_1(0)$, $f_1(1)$ and $f_1(2)$ are defined in (18) and, as we have seen before, allow for tractable closed-form expressions. Moreover, we require the interpolation approximation to satisfy the HT limit as derived in Theorem 5.1:

4. $\lim_{\hat{\rho}\uparrow 1} \mathbb{E}[(1-\hat{\rho})L_{1,app}^{\text{IP}}] = \lim_{\hat{\rho}\uparrow 1} \mathbb{E}[(1-\hat{\rho})L_1] = 1 + \frac{\mu_1 \sigma_C^2}{2m_C}$.

We adhere to the form of (22) with $k = 3$. Then, the four constraints above fully determine the following approximation.

Approximation 5.2. *In the two-layered model, an accurate approximation for the mean queue length of Q_i, based on an interpolation between LT and HT limits, is given by*

$$\mathbb{E}[L_{1,app}^{\text{IP}}] = \frac{a\hat{\rho} + b\hat{\rho}^2 + c\hat{\rho}^3}{1-\hat{\rho}}, \tag{24}$$

where $a = f_1(1)$, $b = f_1(2) - f_1(1)$, $c = 1 + \frac{\mu_1 \sigma_C^2}{2m_C} - f_1(2)$,

$$m_C = \lim_{t\to\infty} \frac{\mathbb{E}[C(t)]}{t} = \boldsymbol{\pi}^{\boldsymbol{\Phi}}_{(U,U)} + \boldsymbol{\pi}^{\boldsymbol{\Phi}}_{(U,R)} \qquad \textit{and} \qquad \sigma_C^2 = \lim_{t\to\infty} \frac{\text{Var}[C(t)]}{t}.$$

We still need to compute σ_C^2. A formal expression for $\text{Var}[C(t)]$ is given in Section 2. However, it is hard to obtain an exact, closed-form expression for σ_C^2. Therefore, we sketch an approach to obtain this value numerically. We approximate σ_C^2 by numerically computing

$$\sigma_C^2 \approx \frac{\mathbb{E}[C^2(2^n)] - (\mathbb{E}[C(2^n)])^2}{2^n} \tag{25}$$

for a large value of n. The computation of the first moment $\mathbb{E}[C(t)]$ is simple and feasible in closed form: $\mathbb{E}[C(t)] = m_C t$. The second moment $\mathbb{E}[C^2(t)]$ can theoretically be obtained by computing $Z(s,t)$ as given in (2), differentiating twice with respect to s and conditioning as desired. However, in practice it is hard to obtain an exact, closed-form expression for $Z(s,t)$. Although numerical values for $\sigma_1, \sigma_2, \nu_1$ and ν_2 are known, s remains a variable. Therefore, we choose to approximate $Z(s,t) = e^{tF(s)}$ by truncating its Taylor series after k terms:

$$Z(s,t) \approx \sum_{i=0}^{k-1} \frac{t^i F^i(s)}{i!}. \tag{26}$$

When taking t large, the number k of terms needed becomes prohibitively large to obtain fairly accurate approximations of $Z(s,t)$. However, for $t = 1$, truncation after $k = 21$ terms generally already produces a very accurate result for $Z(s,1)$. From this, we derive:

$$\mathbb{E}_{i,j}[C(1)] = \frac{-\frac{d}{ds} Z_{i,j}(s,1)}{P_{i,j}(1)} \qquad \text{and} \qquad \mathbb{E}_{i,j}[C^2(1)] = \frac{\frac{d^2}{ds^2} Z_{i,j}(s,1)}{P_{i,j}(1)} \tag{27}$$

for all $i,j \in \{1,\ldots,|\mathcal{S}|\}$, where $P_{i,j}(t)$ and $\mathbb{E}_{i,j}[C(t)]$ are short-hand notations for $\mathbb{P}(\boldsymbol{\Phi}(t) = j | \boldsymbol{\Phi}(0) = i)$ and $\mathbb{E}[C(t)|\boldsymbol{\Phi}(0) = i, \boldsymbol{\Phi}(t) = j]$. The probabilities $P_{i,j}(t)$ are obtained by performing transient analysis on the Markov process $\{\boldsymbol{\Phi}(t) : t \geq 0\}$, or by simply computing $Z_{i,j}(0,t)$. Now that we know how to compute $\mathbb{E}_{i,j}[C(1)]$ and $\mathbb{E}_{i,j}[C^2(1)]$ arbitrarily accurately, we can compute $\mathbb{E}_{i,j}[C(t)]$ and $\mathbb{E}_{i,j}[C^2(t)]$ through a recursion. Under the assumption that the Markov process is already in stationarity at $t = 0$, we obviously have that for $0 < s < t$, $C(t) - C(s)$ is independent of $C(s)$ and has the same distribution as $C(t-s)$. Therefore, the following recursion holds for $t > 0$ by standard probabilistic arguments:

$$\mathbb{E}_{i,j}[C(2t)] = \sum_{k\in\mathcal{S}} \frac{P_{i,k}(t)P_{k,j}(t)}{P_{i,j}(2t)} \Big(\mathbb{E}_{i,k}[C(t)] + \mathbb{E}_{k,j}[C(t)] \Big),$$

$$\mathbb{E}_{i,j}[C^2(2t)] = \sum_{k\in\mathcal{S}} \frac{P_{i,k}(t)P_{k,j}(t)}{P_{i,j}(2t)} \Big(\mathbb{E}_{i,k}[C^2(t)] + 2\mathbb{E}_{i,k}[C(t)]\mathbb{E}_{k,j}[C(t)] + \mathbb{E}_{k,j}[C^2(t)] \Big).$$

Starting with $t = 1$ by using the values in (27), n of these recursion steps provide a value for $\mathbb{E}_{i,j}[C^2(2^n)]$ for all $i, j \in \mathcal{S}$. By conditioning over i, j, this results in

$$\mathbb{E}[C^2(2^n)] = \sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}} \boldsymbol{\pi}_i^{\boldsymbol{\Phi}} P_{i,j}(2^n) \mathbb{E}_{i,j}[C^2(2^n)].$$

By plugging this expression, together with $\mathbb{E}[C(2^n)] = m_C 2^n$, into (25), we can now numerically compute σ_C^2. As the right-hand side of (25) converges very rapidly as n increases, virtually the only numerical error we make stems from (26), which can be made arbitrarily small.

We end this section by observing that Approximation 5.2 performs extremely well. The results equivalent to Table 2 do not show any substantial errors. Like the closed-form approximation obtained in Section 5.1, the interpolation approximation satisfies the LT limits obtained by means of PSA. However, as opposed to the closed-form approximation, the interpolation approximation is also exact when $\hat{\rho} \uparrow 1$. It is therefore intuitively not surprising that the interpolation approximation performs even better than the closed-form approximation, especially for large values of $\hat{\rho}$. Systems for which the interpolation approximation does show substantial errors typically involve fairly loaded queues ($\hat{\rho} \approx 0.7$), and breakdowns and repairs that occur on a far larger time scale than product arrivals and services. For these systems, numerical methods (including simulation) generally fail to work well. When applying PSA numerically, the power series (11) may not converge. Even if it does, one would still have the problem of noticeable truncation errors. Moreover, the time needed to simulate the queue length up to a very accurate degree becomes prohibitively large. We therefore conclude that the accuracy of the interpolation approximation competes with the precision of numerical methods.

Remark 5.1. We derived Approximations 5.1 and 5.2 for a model with two queues and one repairman. However, similar strategies to those used in this section lead to accurate approximations for models with larger numbers of queues and repairmen. To obtain the light-traffic terms a and b, the implementation of PSA must be adapted, as suggested in Remark 3.4. For the heavy-traffic term, Theorem 5.1 still holds. However, since the cardinality of the auxiliary state space $\mathcal{S}$ obviously increases, the computation of σ_C^2 may be computationally more demanding. Similarly, when relaxing the model to allow for phase-type distributed interarrival times, service times, breakdown times and repair times, we can still apply PSA to obtain LT results, as explained in Remark 3.3. To compute the HT-term, Theorem 5.1 needs to be expanded, but this introduces no extra complexity. However, again, the computation of σ_C^2 may be more demanding.

5.3 Behaviour at asymptotic regimes

We conclude by commenting on the behaviour of Approximation 5.1 and Approximation 5.2 in asymptotic instances of the two-layered model.

Light traffic and heavy traffic. By construction, both the closed-form and the interpolation approximations are exact for systems where Q_1 is lightly loaded; i.e., systems where λ_1 tends to zero. Furthermore, the interpolation approximation is by construction exact for systems where Q_1 is fully saturated; i.e., systems where the normalised workload $\hat{\rho}$ tends to one. The latter property is very desirable from a practical perspective, as one is often interested in cases where the queues are heavily loaded. For example, in manufacturing, one is typically interested in maximising the utilisation of machines, without deteriorating significantly the performance of the system.

No M_1-breakdowns. In the asymptotic case where M_1 never breaks down; i.e. $\sigma_1 = 0$, both the closed-form approximation and the interpolation approximation are exact. When there are no M_1-breakdowns, Q_1 behaves like a regular $M/M/1$ queue. For the $M/M/1$ model, it is known that

$$\mathbb{E}[L_1] = \sum_{n=0}^{\infty} \hat{\rho}^{n+1} = \frac{\hat{\rho}}{1-\hat{\rho}}. \tag{28}$$

Since M_1 never breaks down, we obviously have that $m_C = 1$ and $\sigma_C^2 = 0$. Moreover, we have that $f_1(1)|_{\sigma_1=0} = f_1(2)|_{\sigma_2=0} = 1$. Therefore, it is easy to see that (19), (24) and (28) coincide when there are no M_1-breakdowns.

No M_2-breakdowns or instant M_2-repairs. In the asymptotic case, where M_2 does not require any repair time from the repairman, both approximations become exact. Downtimes of Q_1 then only consist of the actual repair times, and are exponentially (ν_1) distributed. Let the completion time C of a type-1 product be the time between the start of its service period and the moment it leaves the system. It is easily verified that Q_1 in isolation can be modelled as an $M/G/1$ queue with server vacations starting at epochs when the queue becomes empty. We refer to this vacation queue as Y. We obtain the expected queue length of Q_1 in this limited regime by studying the mean queue length $\mathbb{E}[L_Y]$ of the equivalent vacation queue Y. The service times in Y correspond to the completion times in Q_1 and the vacation times in Y are composed of the idle times of M_1, plus the downtimes corresponding to breakdowns, which occurred when there was no product in Q_1. Due to the Fuhrmann-Cooper decomposition property [11] applied on Y, the mean queue length of Y can be decomposed as follows:

$$\mathbb{E}[L_Y] = \mathbb{E}[L_{M/G/1}] + \mathbb{E}[L_Y | Y \text{ in vacation period}]. \tag{29}$$

The former term $\mathbb{E}[L_{M/G/1}]$ corresponds to the mean queue length in an $M/G/1$ queue similar to Y, but without any server vacations. The latter term $\mathbb{E}[L_Y|Y$ in vacation period$]$ is the mean queue length in Y observed at a point in time where the server is on vacation. Obviously, this equals the mean number of Poisson (λ_1) arrivals during a residual of a downtime D_1. Since D_1 is exponentially (ν_1) distributed,

$$\mathbb{E}[L_Y|Y \text{ in vacation period}] = \frac{\lambda_1}{\nu_1}.$$

Moreover, it is well-known that

$$\mathbb{E}[L_{M/G/1}] = \lambda_1 \mathbb{E}[C] + \frac{\lambda_1 \mathbb{E}[C^2]}{2(1 - \lambda_1 \mathbb{E}[C])},$$

where the moments $\mathbb{E}[C]$ and $\mathbb{E}[C^2]$ of the completion time can be determined through the relation $C = B_1 + \sum_{i=1}^{N} V_i$. The random variable N is the (geometric) number of repairs needed within a completion time. The repair times V_i are now exponentially (ν_1) distributed. This relation leads to the following Laplace-Stieltjes transform of the completion time:

$$\mathbb{E}[e^{-sC}] = \mathbb{E}[e^{-\left(s+\sigma(1-\mathbb{E}[e^{-sV_1}])\right)B_1}] = \frac{\mu_1}{\mu_1 + s + \sigma(1 - \frac{\nu_1}{\nu_1+s})},$$

out of which the moments of C follow by differentiation with respect to s:

$$\mathbb{E}[C] = \frac{\nu_1 + \sigma_1}{\mu_1 \nu_1} \qquad \text{and} \qquad \mathbb{E}[C^2] = \frac{2\left(\mu_1\sigma_1 + (\nu_1 + \sigma_1)^{\,2}\right)}{\mu_1^2 \nu_1^2}.$$

Since M_2 requires no repair time, we have that $m_C = \frac{\nu_1}{\sigma_1+\nu_1}$ and $\hat{\rho} = \frac{\lambda_1}{\mu_1}\frac{\sigma_1+\nu_1}{\nu_1}$. By combining the results above,

$$\mathbb{E}[L_1] = \frac{\left(1 + \frac{\sigma_1\mu_1}{(\sigma_1+\nu_1)^2}\right)\hat{\rho}}{1 - \hat{\rho}}. \tag{30}$$

For the PSA coefficients, we have that $f_1(1)|_{\sigma_2=0} = f_1(2)|_{\sigma_2=0} = f_1(1)|_{\nu_2\uparrow\infty} = f_1(2)|_{\nu_2\uparrow\infty} = 1 + \frac{\sigma_1\mu_1}{(\sigma_1+\nu_1)^2}$. Since (30) is also exact in HT, Theorem 5.1 implies that $1 + \frac{\mu_1\sigma_C^2}{2m_C} = 1 + \frac{\sigma_1\mu_1}{(\sigma_1+\nu_1)^2}$. Because of these observations, (19), (24) and (30) coincide whenever there are no M_2-breakdowns or M_2-repairs are instant.

Acknowledgements

The authors wish to thank Onno Boxma for valuable comments on earlier drafts of the present paper and Bert Zwart for fruitful discussions on Theorem 5.1.

References

[1] S. Asmussen. *Applied Probability and Queues*. Springer, New York, 2003.

[2] D. Bertsekas and R. Gallager. *Data Networks*. Prentice-Hall, Englewood Cliffs, New Jersey, 1992.

[3] J.P.C. Blanc. A note on waiting times in systems with queues in parallel. *Journal of Applied Probability*, 24:540–546, 1987.

[4] J.P.C. Blanc. On a numerical method for calculating state probabilities for queueuing systems with more than one waiting line. *Journal of Computation and Applied Mathematics*, 20:119–125, 1987.

[5] J.P.C. Blanc. Performance analysis and optimization with the power-series algorithm. In L. Donatiello and R.D. Nelson, editors, *Performance Evaluation of Computer and Communication Systems*, Lecture Notes in Computer Science, pages 53–80. Springer Berlin / Heidelberg, 1993.

[6] M.A.A. Boon, E.M.M. Winands, I.J.B.F. Adan, and A.C.C. van Wijk. Closed-form waiting time approximations for polling systems. *Performance Evaluation*, 68:290–306, 2011.

[7] J.L. Dorsman, R.D. Van der Mei, and E.M.M. Winands. A new method for deriving waiting-time approximations in polling systems with renewal arrivals. *Stochastic Models*, 27:318–332, 2011.

[8] J.L. Dorsman, M. Vlasiou, and O.J. Boxma. Marginal queue length approximations for a two-layered network with correlated queues. Technical Report 2011-43, Eurandom Preprint Series, 2011.

[9] P. J. Fleming and B. Simon. Interpolation approximations of sojourn time distributions. *Operations Research*, 39:251–260, 1991.

[10] G. Franks, T. Al-Omari, M. Woodside, O. Das, and S. Derisavi. Enhanced modeling and solution of layered queuing networks. *IEEE Transactions on Software Engineering*, 35:148–161, 2009.

[11] S.W. Fuhrmann and R.B. Cooper. Stochastic decompositions in the M/G/1 queue with generalized vacations. *Operations Research*, 33:1117–1129, 1985.

[12] D. Gross and J.F. Ince. The machine repair problem with heterogeneous populations. *Operations Research*, 29:532–549, 1981.

[13] M. Harkema, B.M.M. Gijsen, R.D. Van der Mei, and Y. Hoekstra. Middleware performance: A quantitative modeling approach. In *Proceedings of the International Symposium on Performance Evaluation of Computer and Communication Systems (SPECTS)*, pages 733–742, 2004.

[14] G. Hooghiemstra, M. Keane, and S. Van de Ree. Power series for stationary distributions of coupled processor models. *SIAM Journal on Applied Mathematics*, 48:1159–1166, 1988.

[15] J. Keilson and L. D. Servi. The distributional form of Little's law and the Fuhrmann-Cooper decomposition. *Operations Research Letters*, 9:239–247, 1990.

[16] L. Kleinrock. *Queueing Systems, Volume II: Computer Applications*. Wiley, New York, 1976.

[17] S.S. Lavenberg and M. Reiser. Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers. *Journal of Applied Probability*, 17:1048–1061, 1980.

[18] M. I. Reiman and B. Simon. An interpolation approximation for queueing systems with Poisson input. *Operations Research*, 36:454–469, 1988.

[19] L. Takács. *Introduction to the Theory of Queues*. Oxford University Press, New York, 1962.

[20] W.B. Van den Hout and J.P.C. Blanc. Development and justification of the power-series algorithm for BMAP-systems. *Communications in Statistics. Stochastic Models*, 11:471–496, 1995.

[21] P. Wartenhorst. N parallel queueing systems with server breakdown and repair. *European Journal of Operational Research*, 82:302–322, 1995.

[22] W. Whitt. An interpolation approximation for the mean workload in a GI/G/1 queue. *Operations Research*, 37:936–952, 1989.

[23] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, New York, 2002.