Marginal queue length approximations for a two-layered network with correlated queues

EURANDOM PREPRINT SERIES
2011-043 December 14, 2011

Marginal queue length approximations for a two-layered network with correlated queues

J.L. Dorsman, M. Vlasiou, O.J. Boxma
ISSN 1389-2355

Marginal queue length approximations for a two-layered network with correlated queues

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

M. Vlasiou [*] [†]
m.vlasiou@tue.nl

O.J. Boxma [*]
o.j.boxma@tue.nl

December 14, 2011

Abstract

We consider an extension of the classical machine-repair model. As opposed to the classical model, we assume that the machines, apart from receiving service from the repairman, also supply service themselves to queues of products. The extended model can be viewed as a layered queueing network (LQN), where the first layer consists of two separate queues of products. Each of these queues is served by its own machine. The second layer consists of a waiting buffer and a repairman, able to restore the machines into an operational state. When a machine breaks down, it waits in the repair buffer for the repairman to become available. Since the repair time of one machine may affect the period of time the other machine is not able to process products, the downtimes of the machines are correlated. We explicitly model the correlation between the downtimes, which leads to correlation between the queues of products in the first layer. Taking these correlations into account, we obtain approximations for the marginal distributions of the queue lengths in the first layer, by the study of a single server vacation queue. Extensive numerical results show that these approximations are highly accurate.

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. So far, the study of such networks was limited to computer-science problems, see [12] 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. [3]) or as the *time sharing system* (cf. [17, 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 [20, Chapter 5].

We extend this model in two directions. First, we allow machines to have different uptime or repair time distributions. As observed in [15], this leads to technical complications. For example, the arrival theorem (cf. [18])

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.

[*] 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

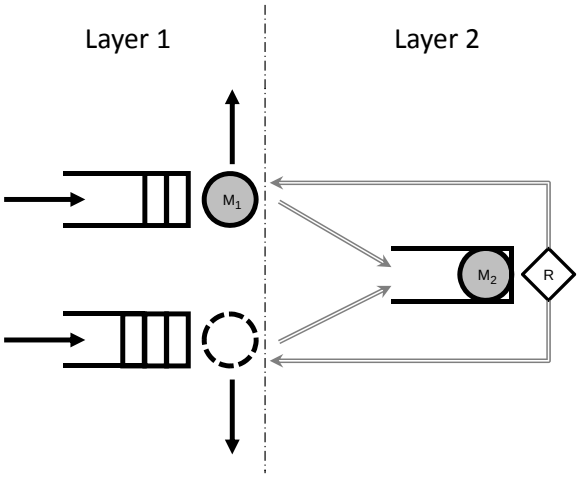

Figure 1: The two-layered model under consideration.

cannot be used any more to derive the stationary downtime distributions of the machines as is done for the original model in [24]. 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 a formulation of a LQN with two layers, which we also refer to as the *two-layered model* or simply the *layered model*.

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. For ease of the discussion, we thus 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. 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. Any progress made is lost, the service requirement resets (*pre-emptive repeat*) and the service starts anew once the machine becomes operational again.

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 upon breakdown the repairman is busy repairing another machine, the machine waits in the buffer. As soon as the other machine is repaired, repair of the current machine starts.

In the present study, we analyse the marginal queue length distributions of the queues in the first layer. 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. This introduces significant positive dependencies in their downtimes and thus in the lengths of the queues in the first layer. If the downtime of one machine is very large, the repair time is probably taking longer than usual, increasing the likelihood for the other machine to break down in the meantime. This has an increasing influence on the next downtime of the first machine, leading to positively correlated consecutive downtimes of a machine. As a result, in the two-layered model, a first-layer queue in isolation can be seen as a single server vacation queue with *dependent* server vacation times. Note that the correlation between the machines' downtimes also leads to correlations in the lengths of the first-layer queues. Hence, there is interaction between the layers, which makes analysis of the first layer rather complicated.

Wartenhorst [24] derives approximations for the first two moments of the queue length distribution of a first-layer queue in a similar model, where he assumes equal uptime and repair time distributions for the machines. In his study, Wartenhorst approximates the first two moments of a first-layer queue with those of a single server vacation queue, where the distribution of the vacations lengths is taken to be equal to that of the machine's downtimes in the layered model, but the vacation lengths are assumed to be *completely independent*. The resulting approximation

is exact by construction for a system where downtimes are independent, and accurate whenever downtimes are only slightly dependent. Since the dependence is completely ignored, Wartenhorst's approximation becomes more inaccurate as the dependence increases. In this paper, we explicitly model the dependence, thus improving accuracy greatly, and obtain an approximation for the *complete distribution* of the queue length.

To approximate the queue length distribution of a given first-layer queue, we draw a connection from such a queue to an M/G/1 queue with server vacations, which we also refer to as the *single server queue*. The M/G/1 queue with server vacations has been studied extensively, see e.g. [8, 9] for surveys. Often vacation lengths or downtimes are assumed to be independent of any other event in the system. Examples of studies where downtimes are assumed to be dependent can be found in [16], where vacation lengths are dependent on the number of customers in the system, and in [6], where vacation lengths are dependent on the length of the previous active period of the server. In the present paper, we assume interdependence between successive vacation lengths, in order to draw the desired connection. In the context of polling systems, vacation queues with interdependent vacation lengths are considered in [1, 10, 14]. However, in that context, the start of a server vacation is usually confined to a point in time where the server concludes the service of a customer. This is not the case in the current context, where a machine can break down at any point in time. For the two-layered model as considered in this paper, the exact correlation structure is not known. Therefore, for the single server queue, we use an explicit, generic dependence form that seems to capture the correlation structure of the problem well.

Section 2 introduces the notation required for both the two-layered model and the single server queueing model, and describes the dependence form used. After this, we analyse the queue length distribution of the single server queue in the latter model at various time epochs through the study of probability generating functions (PGFs). This results in an expression for the (PGF of the) steady state queue length distribution at an arbitrary point in time, which is derived in Section 3. We believe this result to be of independent interest, but our main goal is to apply this result to the study of the layered model. Provided that the dependence in the downtimes of the server in the M/G/1 queue closely matches the dependence of the downtimes of a machine in the layered model, the obtained PGF provides an approximation for the marginal queue length distribution of the corresponding first-layer queue in the layered model. This forms the main result of this paper and is discussed in Section 4. This discussion mainly involves the approximation applied on a two-layered model with two machines and one repairman. However, the same approach can be used to obtain an approximation in models with larger numbers of machines and repairmen. Finally, extensive numerical results from Section 5 show that the obtained approximation is highly accurate, and we identify the factors determining the level of accuracy.

2 Notation

In this section, we introduce the notation used in this paper. Apart from the analysis of the two-layered model, we also derive results on a single server vacation model with dependent vacations. Although notations of the two-layered model and the single server model are mostly equivalent, the two models each define the downtimes, including their dependence, in their own way. Therefore, apart from the two-layered model, we introduce the notation of the single server model separately, including a discussion of the dependence structure assumed.

The two-layered model. The layered model consists of two machines M_1 and M_2 and one repairman R, see Figure 1. Each machine M_i serves its own first-layer queue Q_i on a first-come-first-serve (FCFS) basis. Products arrive at Q_i according to a Poisson process with rate λ_i. The service times required by the products in Q_i are generally distributed according to B_i. The Laplace-Stieltjes transform (LST) of the service time, $\mathbb{E}[e^{-sB_i}]$ ($Re(s) \geq 0$), is denoted by $\widetilde{B}_i(s)$. The steady-state queue length of Q_i, including the product in service, is denoted by L_i. After an exponentially (σ_i) distributed uptime or lifetime, denoted by U_i, a machine M_i will break down, and the service of Q_i inevitably stops. The service of a product in progress is then aborted, and will be restarted once the machine is operational again (*pre-emptive repeat*). When a machine breaks down, it moves to the repair buffer, where it will wait if the repairman is busy repairing the other machine, otherwise the repair will start immediately. A downtime of a machine thus consists of a repair time and possibly a waiting time. The time R_i needed for a repairman to return M_i to an operational state is generally distributed. After a repair, the machine

returns to Q_i and commences service again. Finally, analogous to the notion of load defined for the single server model, we define

$$\rho_i := \lambda_i \mathbb{E}[B_i] < \frac{\mathbb{E}[U_i]}{\mathbb{E}[U_i] + \mathbb{E}[D_i]}. \tag{1}$$

As noted before, the consecutive downtimes are cross-correlated due to the nature of the model. As a result, from the point of view of the products, the machine can be seen as the server in an M/G/1 queue with one-dependent server vacations, where the products themselves are the customers.

Single server model with one-dependent server vacations. In the single server model, the queue is fed by a Poisson process with parameter λ. The service time B required by arriving customers is generally distributed. The uptime U from the moment a server has just ended a vacation period until the start of the next one is exponentially distributed with parameter σ. After this time period U, the server starts a vacation for D time units (a downtime). If a job is in service when the server breaks down, all of the work done on the job is lost and processing of the job is restarted once the server ends its vacation (*pre-emptive repeat*). The steady-state queue length of the queue, including the job in service, is denoted by L. The LST of the service time, $\mathbb{E}[e^{-sB}]$ ($Re(s) \geq 0$), is denoted by $\widetilde{B}(s)$. Likewise, $\widetilde{D}(s)$ represents the LST of D.

The single server model considered differs from most vacation queues studied in literature, by the fact that the durations of vacations (or breakdowns) are one-dependent. For the dependence, we assume a very generic structure, which can be used to model positive correlations between consecutive downtimes. Note that in the two-layered model, consecutive downtimes of a machine are also positively correlated, which motivated our choice. We describe the dependence structure of the downtimes by specifying the LST of a downtime $D(k+1)$ conditioned on its previous downtime $D(k)$:

$$\mathbb{E}[e^{-sD(k+1)}|D(k) = t] = \chi(s)e^{-g(s)t}, \qquad Re(s) \geq 0, \tag{2}$$

where $\chi(s)$ and $g(s)$ are analytic functions in s with $\chi(0) = 1 - g(0) = 1$. Moreover, we assume that $g(s)$ has a completely monotone derivative, i.e. $(-1)^{n+1}\frac{d^n}{ds^n}g(s) \geq 0$ for all $n \geq 1$. Therefore, we have that $e^{-g(s)}$ is the LST of an infinitely divisible distribution (see [11, p. 450]). This generic dependence structure is introduced in [5] to model positive correlation between two random variables. To understand the class of dependence structures (2) describes, let $\{Y(u), u \geq 0\}$ be a stochastic process with stationary independent increments. Since in [11, p. 303] it is shown that the class of distributions of increments in processes with stationary independent increments is identical to the class of infinitely divisible distributions, this means that $Y(t+s) - Y(s)$ is infinitely divisibly distributed and independent of s, for all $s, t > 0$. Let $\chi(s)$ and $e^{-g(s)}$ be the LSTs of $Y(0)$ and $Y(1) - Y(0)$ respectively, where $\chi(s)$ and $e^{-g(s)}$ are assumed to represent LSTs of *positive* continuous random variables. Then, we have that

$$\mathbb{E}[e^{-sY(t)}] = \mathbb{E}[e^{-sY(0)}]\mathbb{E}[e^{-s(Y(t)-Y(0))}] = \mathbb{E}[e^{-sY(0)}]\mathbb{E}[e^{-s(Y(1)-Y(0))}]^t = \chi(s)e^{-g(s)t}.$$

Hence, (2) represents a dependence structure such that $D(k+1) \stackrel{d}{=} Y(t)$ for a certain instance of the process $\{Y(u), u \geq 0\}$, where $t = D(k)$. Thus, $D(k+1)$ consists of two parts, an independent component with LST $\chi(s)$, and a component dependent on the previous downtime, represented by $e^{-g(s)t}$. The class of dependence structures satisfying (2) captures several classic dependence structures, such as linear dependence, compound Poisson processes with rate $\lambda(t)$ and Brownian motions with mean μt and variance $\sigma^2 t$. Note that under the above assumptions, the structure of the downtimes can also be written as a stochastic recursion

$$D(k+1) = G(k)D(k) + X(k),$$

where the values for $G(k)$ and $X(k)$ are independent samples from the distributions represented by the LSTs $e^{-g(s)}$ and $\chi(s)$ respectively. An extensive discussion of this stochastic difference equation is given in e.g. [23].

In the layered model, a downtime can be thought of as a sum of an independent component (e.g., the repair time) and a component dependent on the previous downtime (the waiting time). Therefore, the functions $\chi(s)$ and $g(s)$ can be chosen in such a way that they match the dependence of these downtimes closely.

Note that the functions $\chi(s)$ and $g(s)$ determine the stationary downtime $D := \lim_{k\to\infty} D(k)$. As in steady-state ($k \to \infty$) it holds that $\mathbb{E}[e^{-sD(k+1)}] = \mathbb{E}[e^{-sD(k)}] = \widetilde{D}(s)$, we have that

$$\widetilde{D}(s) = \int_{t=0}^{\infty} \chi(s) e^{-g(s)t} d\mathbb{P}(D < t) = \chi(s)\widetilde{D}(g(s)), \tag{3}$$

giving rise to

$$\mathbb{E}[D] = -\widetilde{D}'(0) = \frac{\chi'(0)}{g'(0) - 1} \text{ and } \mathbb{E}[D^2] = \widetilde{D}''(0) = \frac{\chi''(0) - \mathbb{E}[D](2\chi'(0)g'(0) + g''(0))}{1 - g'(0)^2}. \tag{4}$$

By iteration of (2), one obtains a direct expression for $\widetilde{D}(s)$:

$$\widetilde{D}(s) = \prod_{j=0}^{\infty} \chi(g^{(j)}(s)), \tag{5}$$

where $g^{(0)}(s) = s$ and $g^{(j)}(s) = g(g^{(j-1)}(s))$. The bivariate LST of $D(k)$ and $D(k+1)$ is given by

$$\mathbb{E}[e^{-sD(k)-zD(k+1)}] = \int_{t=0}^{\infty} e^{-st}\mathbb{E}[e^{-zD(k+1)}|D(k) = t]dP(D(k) < t) = \chi(z)\mathbb{E}[e^{-(s+g(z))D(k)}], \tag{6}$$

out of which the joint expectation of two subsequent downtimes $D(k)$ and $D(k+1)$ can be derived:

$$\mathbb{E}[D(k)D(k+1)] = \frac{\partial}{\partial s}\frac{\partial}{\partial z}\chi(z)\mathbb{E}[e^{-(s+g(z))D(k)}]|_{s=0,z=0} = -\chi'(0)\mathbb{E}[D(k)] + g'(0)\mathbb{E}[D(k)^2]. \tag{7}$$

For the steady-state case (i.e. $k \to \infty$), we can obtain a direct expression for the bivariate LST by combining (5) and (6):

$$\lim_{k\to\infty} \mathbb{E}[e^{-sD(k)-zD(k+1)}] = \chi(z) \prod_{j=0}^{\infty} \chi(g^{(j)}(s + g(z))).$$

Finally, the stability condition for the single server model is given by

$$\rho := \lambda\mathbb{E}[B] < \frac{\mathbb{E}[U]}{\mathbb{E}[U] + \mathbb{E}[D]}. \tag{8}$$

3 Analysis of the single server model

Before we can study the length of Q_1 and Q_2 in the layered model, we first perform the analysis required on the single server vacation model. We first derive an expression for the PGF of N, namely the queue length distribution at the beginning of an uptime. For this, we study the transient behaviour of the queue for a duration of two server up-down cycles. An observation length of one cycle would not suffice, since we explicitly need to take the dependence between consecutive downtimes (and thus dependence between cycle lengths) into account. Observe the system in its k^{th} uptime $U(k)$, as well as the following k^{th} downtime $D(k)$ and in the periods $U(k+1)$ and $D(k+1)$ thereafter. Referring to the queue length distribution at the end of an uptime as M, let $N(k)$, $M(k)$, $N(k+1)$, $M(k+1)$ be the corresponding queue lengths, see Figure 2. For $k \to \infty$, we obviously have that

$$\mathbb{E}[p^{N(k)}] = \mathbb{E}[p^{N(k+2)}] = \mathbb{E}[p^N]. \tag{9}$$

In Section 3.1, we first express $\mathbb{E}[p^{N(k+2)}]$ in terms of $\mathbb{E}[p^{N(k)}]$. We do so by observing all the uptimes and downtimes mentioned (see also Figure 2). This leads to an expression for $\mathbb{E}[p^{M(k)}]$ in $\mathbb{E}[p^{N(k)}]$, $\mathbb{E}[p^{N(k+1)}]$ in $\mathbb{E}[p^{M(k)}]$, etc. Having obtained the desired expression from these relations, we compute $\mathbb{E}[p^N]$ in Section 3.2. In Section 3.3, we subsequently compute $\mathbb{E}[p^M]$ and $\mathbb{E}[p^L]$, the PGF of the queue length at an arbitrary point in time. We conclude the analysis of the single server model in Section 3.4 by illustrating the effects of dependence in downtimes. We believe that the analysis of such a single server queue with dependence between successive vacations is not only useful for studying the layered model, but is also of independent interest.

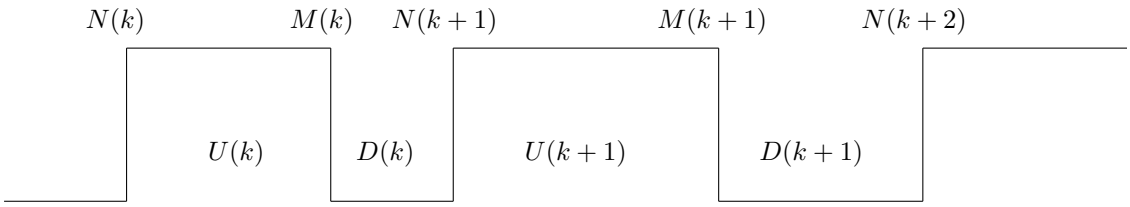

Figure 2: Two server up/down cycles.

3.1 Behaviour of the queue length in two server up/down cycles

To obtain a relation between $\mathbb{E}[p^{N(k+2)}]$ and $\mathbb{E}[p^{N(k)}]$, we observe the way the queue length evolves in each of the periods $U(k)$, $D(k)$, $U(k+1)$ and $D(k+1)$. Connecting the results leads to an expression for $\mathbb{E}[p^{N(k+2)}]$ in terms of $\mathbb{E}[p^{N(k)}]$.

3.1.1 The queue length distribution during the first uptime

In this section, we derive a relation between $\mathbb{E}[p^{M(k)}]$ and $\mathbb{E}[p^{N(k)}]$. During the first uptime $U(k)$, the server is accepting and processing customers. This means that the queue length in this period of time evolves similarly to the length of a regular M/G/1 queue during an exponential (σ) interval. This M/G/1 queue has the same customer arrival process and the same service time distribution, but does not have any service interruptions or server downtimes.

A relation between the PGFs of the queue length distribution at the beginning and the end of an exponentially distributed time interval in an M/G/1 queue can be obtained from the queue length transition probabilities between these two points in time. In e.g. [7, p. 246], these transition probabilities are derived as well as the resulting relation between the queue lengths at the beginning and end of an exponentially distributed time interval. The relation between $M(k)$ and $N(k)$ in our context immediately follows:

$$\mathbb{E}[p^{M(k)}] = A(p) * \mathbb{E}[p^{N(k)}] + K(p) * \mathbb{E}[\mu^{N(k)}(\sigma)], \tag{10}$$

where

$$\begin{aligned} A(p) &= \frac{\sigma}{\sigma+\lambda(1-p)} \frac{p(1-\widetilde{B}(\sigma+\lambda(1-p)))}{p-\widetilde{B}(\sigma+\lambda(1-p))}, \\ K(p) &= -\frac{\sigma}{\sigma+\lambda(1-\mu(\sigma))} \frac{(1-p)\widetilde{B}(\sigma+\lambda(1-p))}{p-\widetilde{B}(\sigma+\lambda(1-p))} \end{aligned}$$

and $\mu(\sigma)$ is the LST of a busy period in the regular M/G/1 queue evaluated at σ. The value $\mu(\sigma)$ is the unique root of the expression $p-\widetilde{B}(\sigma+\lambda(1-p))$ with $|\mu(\sigma)|<1$ (for a proof of uniqueness, see [20, p. 47–49]). Therefore, $\mu(\sigma)$ is a pole of both $A(p)$ and $K(p)$, but these poles compensate each other. More specifically, we find by standard methods the following result that we will need in the sequel:

$$\begin{aligned} &\lim_{p\to\mu(\sigma)} \big[A(p)+K(p)\big] \\ &\quad = \lim_{p\to\mu(\sigma)} \Big(\frac{\sigma}{\sigma+\lambda(1-p)} + \Big(\frac{\sigma}{\sigma+\lambda(1-p)} - \frac{\sigma}{\sigma+\lambda(1-\mu(\sigma))} \Big) \frac{(1-p)\widetilde{B}(\sigma+\lambda(1-p))}{p-\widetilde{B}(\sigma+\lambda(1-p))} \Big) \\ &\quad = \frac{\sigma}{\sigma+\lambda(1-\mu(\sigma))} + \frac{\lambda\mu(\sigma)\sigma(1-\mu(\sigma))}{\big(1+\lambda B'(\sigma+\lambda(1-\mu(\sigma)))\big)\big(\sigma+\lambda(1-\mu(\sigma))\big)^2}. \end{aligned} \tag{11}$$

3.1.2 The queue length distribution during the first downtime

During the first downtime $D(k)$, the server does not process any customers. Therefore, the queue length increases by the number of customer arrivals in this period. More specifically, the difference between $M(k)$ and $N(k+1)$ is exactly the number of Poisson arrivals during $D(k)$. It will prove convenient in later calculations to condition on the event $D(k) = t$ for any $t \in \mathbb{R}_+$. Let $H(t)$ be Poisson (λt) distributed, i.e. the number of Poisson arrivals during $D(k) = t$. We then obtain the following relation between $\mathbb{E}[p^{N(k+1)}|D(k) = t]$ and $\mathbb{E}[p^{M(k)}]$:

$$\begin{aligned}\mathbb{E}[p^{N(k+1)}|D(k) = t] =& \mathbb{E}[p^{M(k)+H(t)}] \\ =& \mathbb{E}[p^{M(k)}] \sum_{i=0}^{\infty} p^i e^{-\lambda t} \frac{(\lambda t)^i}{i!} \\ =& \mathbb{E}[p^{M(k)}] e^{-\lambda(1-p)t}. \end{aligned} \tag{12}$$

3.1.3 The queue length distribution during the second uptime

In this section, we obtain a relation between $\mathbb{E}[p^{M(k+1)}|D(k) = t]$ and $\mathbb{E}[p^{N(k+1)}|D(k) = t]$. During the second uptime $U(k+1)$, the server is processing customers for an exponentially (σ) distributed amount of time, which means that the analysis is largely the same as the analysis of the queue length during the first uptime $U(k)$. The only difference stems from the fact that we now choose to condition on the event $D(k) = t$, in order to be able to concatenate all the results later on. Analogous to (10), we have

$$\mathbb{E}[p^{M(k+1)}|D(k) = t] = A(p) * \mathbb{E}[p^{N(k+1)}|D(k) = t] + K(p) * \mathbb{E}[\mu^{N(k+1)}(\sigma)|D(k) = t] \tag{13}$$

with $A(p)$, $K(p)$ and $\mu(\sigma)$ as before.

3.1.4 The queue length distribution during the second downtime

To obtain a relation between $\mathbb{E}[p^{N(k+2)}|D(k) = t]$ and $\mathbb{E}[p^{M(k+1)}|D(k) = t]$, note that the server is not processing customers during the period $D(k+1)$, which again means that the difference between $M(k+1)$ and $N(k+2)$ is equal to the number of Poisson arrivals during the period $D(k+1)$. The duration of $D(k+1)$ is dependent on $D(k)$, which is described by the LST in (2) conditioned on $D(k) = t$. Therefore, the previously introduced conditioning on the event $D(k) = t$ for $t \in \mathbb{R}$ is convenient at this point. Extending the analysis of Section 3.1.2 to the second downtime, conditioned on the duration of the first downtime, we implement the dependence in (2) and obtain the following relation:

$$\begin{aligned}\mathbb{E}[p^{N(k+2)}|D(k) = t] &= \int_{u=0}^{\infty} \mathbb{E}[p^{M(k+1)+H(u)}|D(k) = t] \, d\mathbb{P}(D(k+1) < u|D(k) = t) \\ &= \int_{u=0}^{\infty} \mathbb{E}[p^{M(k+1)}|D(k) = t] e^{-\lambda(1-p)u} \, d\mathbb{P}(D(k+1) < u|D(k) = t) \\ &= \mathbb{E}[p^{M(k+1)}|D(k) = t] \mathbb{E}[e^{-\lambda(1-p)D(k+1)}|D(k) = t] \\ &= \mathbb{E}[p^{M(k+1)}|D(k) = t] \chi(\lambda(1-p)) e^{-g(\lambda(1-p))t}, \end{aligned} \tag{14}$$

where $\mathbb{E}[p^{H(u)}] = e^{-\lambda(1-p)u}$ is the PGF of the number of Poisson arrivals during a time period u.

3.1.5 Connecting all periods

Combining (10), (12), (13) and (14), we obtain an expression for $\mathbb{E}[p^{N(k+2)}|D(k) = t]$ in terms of $\mathbb{E}[p^{N(k)}]$. Keeping in mind (11) and the fact that $\mu(\sigma)$ is a pole of $A(p)$ and $K(p)$, we note that for the substitution of

$\mathbb{E}[\mu^{M(k)}(\sigma)]$ the following important observation holds:

$$\begin{aligned}
\lim_{p\to\mu(\sigma)}\mathbb{E}[p^{M(k)}] &= \lim_{p\to\mu(\sigma)}\sum_{i=0}^{\infty}\big(A(p)*p^i+K(p)*\mu^i(\sigma)\big)\mathbb{P}(N(k)=i)\\
&=\sum_{i=0}^{\infty}\Big(\lim_{p\to\mu(\sigma)}\big[A(p)+K(p)\big]\mu^i(\sigma)+\lim_{p\to\mu(\sigma)}\big[A(p)(p^i-\mu^i(\sigma))\big]\Big)\mathbb{P}(N(k)=i)\\
&=\sum_{i=0}^{\infty}\Big(\lim_{p\to\mu(\sigma)}\big[A(p)+K(p)\big]\mu^i(\sigma)\\
&\qquad\qquad+\frac{\sigma(1-\mu(\sigma))}{(\sigma+\lambda(1-\mu(\sigma)))(1+\lambda\widetilde{B}'(\sigma+\lambda(1-\mu(\sigma))))}i\mu^i\Big)\mathbb{P}(N(k)=i)\\
&=\Big(\lim_{p\to\mu(\sigma)}\big[A(p)+K(p)\big]\Big)\mathbb{E}[\mu^{N(k)}(\sigma)]\\
&\qquad\qquad+\frac{\sigma(1-\mu(\sigma))}{(\sigma+\lambda(1-\mu(\sigma)))(1+\lambda\widetilde{B}'(\sigma+\lambda(1-\mu(\sigma))))}\mathbb{E}[N(k)\mu^{N(k)}(\sigma)].
\end{aligned}$$

Hence, an extra term containing $\mathbb{E}[N(k)\mu^{N(k)}(\sigma)]$ arises in the expression for $\mathbb{E}[p^{N(k+2)}|D(k)=t]$. We obtain

$$\begin{aligned}
&\mathbb{E}[p^{N(k+2)}|D(k)=t]=\chi(\lambda(1-p))A^2(p)e^{-(\lambda(1-p)+g(\lambda(1-p)))t}\mathbb{E}[p^{N(k)}]\\
&+\chi(\lambda(1-p))K(p)\Big(A(p)e^{-(g(\lambda(1-p))+\lambda(1-p))t}\\
&\qquad\qquad+\lim_{p\to\mu(\sigma)}\big[A(p)+K(p)\big]e^{-(g(\lambda(1-p))+\lambda(1-\mu(\sigma)))t}\Big)\mathbb{E}[\mu^{N(k)}(\sigma)]\\
&+\chi(\lambda(1-p))K(p)e^{-(g(\lambda(1-p))+\lambda(1-\mu(\sigma)))t}\\
&\qquad\qquad*\frac{\sigma(1-\mu(\sigma))}{(\sigma+\lambda(1-\mu(\sigma)))(1+\lambda\widetilde{B}'(\sigma+\lambda(1-\mu(\sigma))))}\mathbb{E}[N(k)\mu^{N(k)}(\sigma)].
\end{aligned}\tag{15}$$

In the course of the previous calculations, we conditioned on the event $D(k)=t$ in order to incorporate the dependence between the downtimes. In the expression for $\mathbb{E}[p^{N(k+2)}|D(k)=t]$, we see that the value t is only found in the form $e^{-st}(s\geq 0)$, meaning that unconditioning leads to expressions in terms of the LST $\widetilde{D}(\cdot)$:

$$\begin{aligned}
\mathbb{E}[p^{N(k+2)}] &= \int_{t=0}^{\infty}\mathbb{E}[p^{N(k+2)}|D(k)=t]d\mathbb{P}(D(k)<t)\\
&= E(p)\mathbb{E}[p^{N(k)}]+F(p)\mathbb{E}[\mu^{N(k)}(\sigma)]+G(p)\mathbb{E}[N(k)\mu^{N(k)}(\sigma)],
\end{aligned}\tag{16}$$

with

$$\begin{aligned}
E(p) =&\chi(\lambda(1-p))A^2(p)\widetilde{D}(\lambda(1-p)+g(\lambda(1-p))),\qquad(17)\\
F(p) =&\chi(\lambda(1-p))K(p)\Big(A(p)\widetilde{D}(\lambda(1-p)+g(\lambda(1-p)))\\
&\qquad+\widetilde{D}(\lambda(1-\mu(\sigma))+g(\lambda(1-p)))\lim_{z\to\mu(\sigma)}[A(z)+K(z)]\Big),\\
G(p) =&\chi(\lambda(1-p))K(p)\widetilde{D}(\lambda(1-\mu(\sigma))+g(\lambda(1-p)))\\
&\qquad*\frac{\sigma(1-\mu(\sigma))}{(\sigma+\lambda(1-\mu(\sigma)))(1+\lambda\widetilde{B}(\sigma+\lambda(1-\mu(\sigma))))}.
\end{aligned}$$

This expression gives a relation between $\mathbb{E}[p^{N(k+2)}]$ and $\mathbb{E}[p^{N(k)}]$.

3.2 The queue length distribution at the beginning of an arbitrary uptime

We now compute $\mathbb{E}[p^N]=\lim_{k\to\infty}\mathbb{E}[p^{N(k)}]$. Combining (9) and (16), we find

$$\mathbb{E}[p^N]=\frac{F(p)\mathbb{E}[\mu^N(\sigma)]+G(p)\mathbb{E}[N\mu^N(\sigma)]}{1-E(p)}\tag{18}$$

with $E(p), F(p)$ and $G(p)$ as before. Observe that this expression has two unknown constants $\mathbb{E}[\mu^N(\sigma)]$ and $\mathbb{E}[N\mu^N(\sigma)]$. We show that these constants can be obtained as the solution of a system of two linear equations. First, we obviously have that for $p = 1$ the LHS of (18) equals one, leading to the first equation. Another equation can be obtained by noting that the denominator of (18) has a root ϕ between $p = 0$ and $p = 1$. This implies that, as $\mathbb{E}[p^N]$ is analytic for $|p| < 1$, the numerator should evaluate to zero for $p = \phi$ as well, leading to the second equation. These two linear equations lead to a unique solution for $\mathbb{E}[\mu^N(\sigma)]$ and $\mathbb{E}[N\mu^N(\sigma)]$. We derive these equations below. Expressions for the constants immediately follow.

The case $p = 1$. Since the LHS of (18) evaluates to one for $p = 1$ and $F(1) = G(1) = 1 - E(1) = 0$, we have for the RHS

$$\lim_{p\to 1}\big[\frac{F(p)\mathbb{E}[\mu^N(\sigma)] + G(p)\mathbb{E}[N\mu^N(\sigma)]}{1 - E(p)}\big] = -\frac{F'(1)\mathbb{E}[\mu^N(\sigma)] + G'(1)\mathbb{E}[N\mu^N(\sigma)]}{E'(1)} = 1$$

by l'Hôpital's rule. Since $E(p)$, $F(p)$ and $G(p)$ are each differentiable at $p = 1$, this results in the first linear equation in the two unknowns $\mathbb{E}[\mu^N(\sigma)]$ and $\mathbb{E}[N\mu^N(\sigma)]$.

The case $p = \phi$. The denominator $1 - E(p)$ of (18) has a root $p = \phi$ between zero and $\mu(\sigma) < 1$. More specifically, we have the following:

Lemma 3.1. *The denominator $1 - E(p)$ has exactly one root on the real line in the domain $(0, \mu(\sigma))$.*

Proof. See Appendix A. □

Let ϕ be the unique root mentioned in Lemma 3.1. Since $\mathbb{E}[p^N]$ is analytic in p for $|p| \leq 1$ and thus cannot evaluate to $\pm\infty$ for $0 < p < \mu(\sigma)$, we have that this root should also be a root for the numerator. Hence, we have that $F(\phi)\mathbb{E}[\mu^N(\sigma)] + G(\phi)\mathbb{E}[N\mu^N(\sigma)] = 0$.

Combining (18) with the cases $p = 1$ and $p = \phi$, we obtain the following lemma:

Lemma 3.2. *The PGF of the queue length at the beginning of an arbitrary uptime is given by*

$$\mathbb{E}[p^N] = \frac{F(p)\mathbb{E}[\mu^N(\sigma)] + G(p)\mathbb{E}[N\mu^N(\sigma)]}{1 - E(p)},$$

where

$$\mathbb{E}[\mu^N(\sigma)] = \frac{E'(1)G(\phi)}{F(\phi)G'(1) - F'(1)G(\phi)} \text{ and } \mathbb{E}[N\mu^N(\sigma)] = \frac{E'(1)F(\phi)}{F'(1)G(\phi) - F(\phi)G'(1)}.$$

Remark 3.1. Note that in case additional real roots for the denominator of $\mathbb{E}[p^N]$ exist in the domain $(\mu(\sigma), 1)$ or non-real roots exist in $|p| < 1$, any resulting additional equations cannot contradict the equations found for $\mathbb{E}[\mu^N(\sigma)]$ and $\mathbb{E}[N\mu^N(\sigma)]$. The constants are guaranteed to exist for a stable system, since a unique limiting distribution for N exists as well. Therefore, and because it is technically very involved, we refrain from providing a detailed proof (via Rouché's theorem) that no other roots exist in $|p| < 1$.

3.3 The queue length distribution at an arbitrary point in time

The main goal of this section is to determine the (PGF of the) queue length distribution at an arbitrary point in time. To do so, we expand the results of the previous section. The PGF $\mathbb{E}[p^M]$ of the queue length at the start of an arbitrary downtime is easily derived from the PGF $\mathbb{E}[p^N]$ of the queue length at the start of an arbitrary uptime. We then obtain the PGFs of the queue length when observed at an arbitrary point within an uptime and when observed at an arbitrary point within a downtime respectively. As a result, we finally obtain a general expression for $\mathbb{E}[p^L]$, the PGF of the queue length at an arbitrary point in time.

3.3.1 Observing the queue length during an arbitrary uptime

For the distribution of the queue length at an arbitrary point during an arbitrary uptime, we first obtain an expression for $\mathbb{E}[p^M]$. Then, we show that the PGF of the desired distribution equals this expression.

Following the same reasoning as in Sections 3.1.2 and 3.1.4, we derive $\mathbb{E}[p^M]$ by noting that during a downtime D between M and N, new customers arrive, but no customers are being processed:

$$\begin{aligned}\mathbb{E}[p^N] =& \int_{t=0}^{\infty} \mathbb{E}[p^M]\mathbb{E}[p^{H(t)}]d\mathbb{P}(D < t) \\ =& \mathbb{E}[p^M]\widetilde{D}(\lambda(1-p)),\end{aligned} \tag{19}$$

where $H(t)$ is the number of Poisson (λ) arrivals during a time interval t. Thus, the following lemma can be derived:

Lemma 3.3. *The PGF of the queue length at an arbitrary point in an uptime is given by*

$$\mathbb{E}[p^L|\textit{server up}] = \frac{\mathbb{E}[p^N]}{\widetilde{D}(\lambda(1-p))},$$

where $\mathbb{E}[p^N]$ is given in Lemma 3.2 and $\widetilde{D}(\cdot)$ satisfies (3) *and* (5)*.*

Proof. Let $V(t)$ be the number of vacation initiations of the server in $(0, t]$. Note that $V(t)$ is a doubly stochastic process, where during a server uptime initiations of vacations occur according to a Poisson process with rate σ, whereas they obviously occur with rate zero when the server is already on a vacation. The conditional PASTA property (cf. [22]) applied to $V(t)$ implies that the queue length distribution at the start of vacations equals the queue length distribution at an arbitrary point in time during an uptime. Hence, $\mathbb{E}[p^L|\textit{server up}] = \mathbb{E}[p^M]$. Combining this with (19) yields the result. □

Remark 3.2. An expression for $\mathbb{E}[p^M]$ into $\mathbb{E}[p^N]$ is also readily given by (10). This leads to an alternative expression for the PGF of the queue length when observed during a downtime:

$$\mathbb{E}[p^L|\textit{server up}] = A(p)\mathbb{E}[p^N] + K(p)\mathbb{E}[\mu^N(\sigma)],$$

with $A(p)$, $K(p)$ and $\mu(\sigma)$ as before.

3.3.2 Observing the queue length during a downtime

At an arbitrary point in time during a downtime, the number of customers in the system can be decomposed into the number of customers who were already waiting at the end of the previous uptime M, and the number of customers who arrived during the elapsed time D^{past} since the start of the *current* downtime, which we denote with $H(D^{past})$. Note that M and $H(D^{past})$ are not independent. A large value of M may imply that the *previous* downtime has been very long. Due to the positive correlation between the downtimes as assumed in Section 2, this would in its turn imply that the *current* downtime is probably longer than usual as well. The *current* downtime and its past time D^{past} are obviously dependent, which results in the fact that M and $H(D^{past})$ are dependent.

Using the notation illustrated in Figure 2, we obtain

$$\begin{aligned}\mathbb{E}[p^L|\textit{server down}] &= \mathbb{E}[p^{M+H(D^{past})}] \\ &= \lim_{k\to\infty} \int_0^{\infty} \mathbb{E}[p^{M(k+1)}|D(k) = t]\mathbb{E}[p^{H(D^{past}(k+1))}|D(k) = t]d\mathbb{P}(D(k) < t).\end{aligned} \tag{20}$$

From the intermediate calculations leading to (16) (or by simply combining (14) and (15)), we have that

$$\lim_{k\to\infty} \mathbb{E}[p^{M(k+1)}|D(k) = t] = \sum_{i=1}^{2} q_i(p)e^{-r_i(p)t}, \tag{21}$$

where

$$q_1(p) = A(p)(A(p)\mathbb{E}[p^N] + K(p)\mathbb{E}[\mu^N(\sigma)]), \tag{22}$$

$$q_2(p) = K(p)\Big(\Big(\lim_{z\to\mu(\sigma)}[A(z)+K(z)]\Big)\mathbb{E}[\mu^N(\sigma)] \tag{23}$$

$$+ \frac{\sigma(1-\mu(\sigma))}{(\sigma+\lambda(1-\mu(\sigma)))(1+\lambda\widetilde{B}'(\sigma+\lambda(1-\mu(\sigma))))}\mathbb{E}[N\mu^N(\sigma)]\Big), \tag{24}$$

$$r_1(p) = \lambda(1-p) \text{ and } r_2(p) = \lambda(1-\mu(\sigma)). \tag{25}$$

Moreover, from (2) we obtain

$$\begin{aligned}\lim_{k\to\infty}\mathbb{E}[p^{H(D^{past}(k+1))}|D(k)=t] &= \mathbb{E}[e^{-\lambda(1-p)D^{past}(k+1)}|D(k)=t] \\ &= \frac{1-\mathbb{E}[e^{-\lambda(1-p)D(k+1)}|D(k)=t]}{\lambda(1-p)\mathbb{E}[D(k+1)|D(k)=t]} \\ &= \frac{1-\chi(\lambda(1-p))e^{-g(\lambda(1-p))t}}{\lambda(1-p)(g'(0)t-\chi'(0))}.\end{aligned} \tag{26}$$

Combining (21)-(26), we have that the evaluation of (20) involves the computation of a linear combination of integrals with the form

$$\int_{t=0}^{\infty}\frac{e^{-at}}{bt+c}d\mathbb{P}(D<t) = \int_{t=0}^{\infty}\int_{u=0}^{\infty}e^{-(at+(bt+c)u)}du\,d\mathbb{P}(D<t).$$

By interchanging the integrals, this expression reduces to

$$\int_0^{\infty}e^{-cu}\widetilde{D}(a+bu)du =: \kappa_{[a,b]}\{c\},$$

i.e. the Laplace transform of the function $\gamma_{[a,b]}\{u\} := \widetilde{D}(a+bu)$. We obtain the following lemma:

Lemma 3.4. *The PGF of the queue length at an arbitrary point in a downtime is given by*

$$\begin{aligned}\mathbb{E}[p^L|\textit{server down}] &= \int_{t=0}^{\infty}\sum_{i=1}^{2}q_i(p)e^{-r_i(p)t}\frac{1-\chi(\lambda(1-p))e^{-g(\lambda(1-p))t}}{\lambda(1-p)(g'(0)t-\chi'(0))}d\mathbb{P}(D<t) \\ &= \frac{1}{\lambda(1-p)}\sum_{i=1}^{2}q_i(p)\Big(\kappa_{[r_i(p),g'(0)]}\{-\chi'(0)\} \\ &\qquad - \chi(\lambda(1-p))\kappa_{[r_i(p)+g(\lambda(1-p)),g'(0)]}\{-\chi'(0)\}\Big),\end{aligned} \tag{27}$$

where $\gamma_{[a,b]}\{u\} = \widetilde{D}(a+bu)$ *and* $\kappa_{[a,b]}\{c\} = \int_0^{\infty}e^{-cu}\widetilde{D}(a+bu)du$*, the Laplace transform of* $\gamma_{[a,b]}\{\cdot\}$.

Note that in case $\widetilde{D}(\cdot)$ is not explicitly known by inspecting (3), one can still evaluate $\kappa_{[a,b]}\{c\}$ up to arbitrary precision by truncating the infinite product form in (5).

3.3.3 Deriving the general queue length distribution

Now that we have the PGF of L conditioned on the state of the server, we readily have the queue length distribution of the single server model:

Theorem 3.5. *For the PGF of the queue length in the single server model with one-dependent downtimes, we have*

$$\mathbb{E}[p^L] = p_{up}\mathbb{E}[p^L|\textit{server up}] + p_{down}\mathbb{E}[p^L|\textit{server down}], \tag{28}$$

where

$$p_{up} = \frac{\mathbb{E}[U]}{\mathbb{E}[U]+\mathbb{E}[D]} = \frac{1}{1+\sigma\mathbb{E}[D]} \text{ and } p_{down} = \frac{\mathbb{E}[D]}{\mathbb{E}[U]+\mathbb{E}[D]} = \frac{\sigma\mathbb{E}[D]}{1+\sigma\mathbb{E}[D]}.$$

Expressions for $\mathbb{E}[p^L|\textit{server up}]$ and $\mathbb{E}[p^L|\textit{server down}]$ are given in Lemmas 3.3 and 3.4 respectively. The weights p_{up} and p_{down} are the probabilities that one finds the server up and down respectively when observing the system at a random point in time in steady-state. These probabilities are derived through the straightforward application of Palm theory (cf. [2, 19]) and involve the computation of $\mathbb{E}[U]$ and $\mathbb{E}[D]$. The former is determined by the fact that U is exponentially (σ) distributed, and the latter follows from (4).

The obtained expression for the (PGF of the) queue length distribution is exact for the single server vacation model with server vacation times dependent according to (2). It can serve as an approximation for the marginal queue length distribution of a first-layer queue in the layered model, which is examined in Section 4. We end this subsection with two remarks.

Remark 3.3. Observe that the evaluation of (28) involves the evaluation of several values of the downtime LST $\widetilde{D}(\cdot)$. Whenever the downtime LST is not readily derived by (3), computing the values of $\widetilde{D}(\cdot)$ is not possible in an exact fashion. However, we can use the infinite product form representation (5) to derive these values up to arbitrary precision. This product converges fast and therefore truncation leads to an arbitrarily accurate approximation.

Remark 3.4. The analysis of the single server queue as presented in this section can be extended to other dependence forms than (2). For example, for Markov-modulated dependencies the same strategy can be used to obtain queue length distributions. Slight adaptations have to be made in the computations, starting with the conditional LST term in (14).

3.4 A note on the impact of dependence

Now that we have obtained the PGF of the queue length, we numerically study the influence of the downtime dependence on the queue length distribution. Observe an instance of the single server model where λ = 3, the service time B is exponentially distributed with rate 5, and the uptime U of the server is exponentially distributed with rate 1/3. In this particular example, the downtime of the server consists of multiple exponential phases. The number of phases of which a downtime $D(k+1)$ consists, depends on the previous downtime $D(k)$:

$$D(k+1) \stackrel{d}{=} C_1 + \cdots + C_{J(D(k))+1}, \tag{29}$$

where the C_i are i.i.d. exponentially (δ) distributed, $\delta > 1$, and $J(\mu)$ is Poisson distributed with parameter μ. This implies that

$$\begin{aligned}\mathbb{E}[e^{-sD(k+1)}|D(k)=t] &= \sum_{j=0}^{\infty} \mathbb{E}[e^{-s(C_1+\sum_{i=2}^{j+1} C_i)}]e^{-t}\frac{t^j}{j!} \\ &= \mathbb{E}[e^{-sC_1}]\sum_{j=0}^{\infty}\mathbb{E}[e^{-sC_1}]^j e^{-t}\frac{t^j}{j!} \\ &= \mathbb{E}[e^{-sC_1}]e^{-(1-\mathbb{E}[e^{-sC_1}])t}.\end{aligned}$$

Therefore, we have that $\chi(s) = \mathbb{E}[e^{-sC_1}] = \frac{\delta}{\delta+s}$ and $g(s) = 1-\mathbb{E}[e^{-sC_1}] = \frac{s}{\delta+s}$. The stationary downtime is exponentially ($\delta-1$) distributed, since (3) is satisfied for its LST $\widetilde{D}(s) = \frac{\delta-1}{\delta-1+s}$. Note that the stationary downtime distribution only exists for $\delta > 1$.

We compare the model above with its 'independent counterpart', namely a single server queue with the same interarrival, service, uptime and stationary downtime distributions as before, but with mutually independent downtimes. Note that independent downtimes also fit in the dependence structure of (2) by simply setting $g(s) = 0$

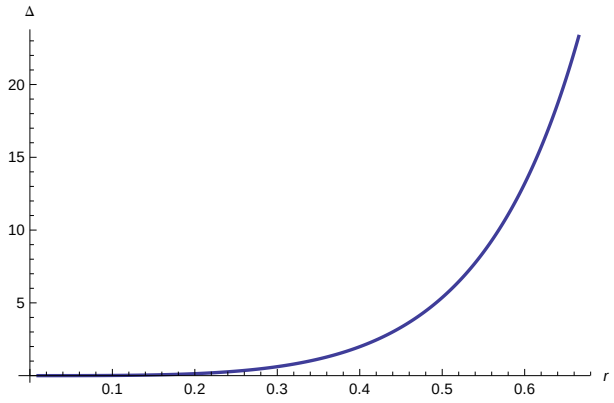

Figure 3: The relative difference Δ in $\mathbb{E}[L]$ between the dependent and the independent model for various values of $\mathbb{E}[D]$.

for all s. Since the stationary downtime distribution is exponentially $(\delta - 1)$ distributed, we trivially have for the independent model that $\chi(s) = \widetilde{D}(s) = \frac{\delta-1}{\delta-1+s}$, $g(s) = 0$.

To see the effect of the dependencies, we compare the expected queue length in the dependent model, $\mathbb{E}[L^{dep}]$, with that of the independent model, $\mathbb{E}[L^{indep}]$. These values are obtained by evaluating the derivative of (28) at $p = 1$. We compute the percentual relative difference of both quantities, i.e.

$$\Delta := 100\% \times \frac{\mathbb{E}[L^{dep}] - \mathbb{E}[L^{indep}]}{\mathbb{E}[L^{indep}]},$$

for varying values of δ such that the load of the system varies between 0.6 and 1. Note that for the dependent model, the value of δ determines the correlation coefficient between two consecutive downtimes in steady-state, which we denote by r. More specifically, we have by the definition of the correlation coefficient that

$$r = \frac{\lim_{k\to\infty} \mathbb{E}[D(k)D(k+1)] - (\mathbb{E}[D])^2}{\mathbb{E}[D^2] - (\mathbb{E}[D])^2}. \tag{30}$$

This expression can be further expressed in terms of δ by using (4) and (7). For the independent model, the correlation coefficient between the downtimes obviously equals zero at all times. Figure 3 shows the value of Δ as a function of the correlation r as observed in the dependent model. We see in this figure that Δ equals zero for $r = 0$, while Δ grows as high as 25% for increasing r. Thus, if we would approximate the mean queue length in the dependent model by ignoring the correlation between the downtimes, we can have an error of 25%. This figure shows that the correlation in the downtimes can have a large impact on the queue length.

4 Approximating queue lengths in the two-layered model

In this section, we draw an analogy between the single server model and the two-layered model. More specifically, we use Theorem 3.5 to obtain approximations for the marginal queue length distributions in the layered model. Notions for arrival streams, service times and uptimes are equivalent for both models. To describe the downtime distribution and the dependence of the downtimes in the layered model in terms of the parameters of the single server model, we need to obtain suitable choices for the functions $\chi(s)$ and $g(s)$, which are used in (2). The quality of the resulting approximation for L_i, the queue length of Q_i, is strongly dependent on these choices. Note that if these two functions modelled the dependence in the two-layered model perfectly, the result on the single server model would also be exact for the two-layered model, rather than merely an accurate approximation.

In Section 4.1, we first compute the first two moments and the correlation coefficient of subsequent downtimes in the layered model under the assumption that the repair times of the machines are exponentially distributed, for

illustrative purposes. This subclass of the model theoretically allows for exact solutions, since a first-layer queue in isolation in this subclass can be modelled as an M/G/1 queue with Markov-modulated server availability. For general repair time distributions, the computation of the correlation coefficient is more complicated. Based on the obtained moments and correlation coefficient, we derive suitable choices for the functions $\chi_i(s)$ and $g_i(s)$ in Section 4.2, such that they match the situation in the layered model as good as possible. After these preliminary steps, we combine these results with those of the previous section to obtain an approximation for (the PGF of) L_i, one of the main results of this paper, in Section 4.3. This approximation is applicable for the layered model with two machines and one single repairmen. Note however that the approach followed remains valid to obtain approximations for more general models. We discuss this in Section 4.4.

4.1 Moments and the correlation coefficient of downtimes in the layered model

In this section, we derive the first two moments of the stationary downtime distribution of machine M_1 in the layered model, as well as the correlation coefficient between two subsequent downtimes $D_1(k)$ and $D_1(k+1)$ in steady-state (i.e., for $k \to \infty$), for the case that the repair times R_1 and R_2 of both machines are exponentially (ν_1) and exponentially (ν_2) distributed respectively. We do this by studying the bivariate LST $\mathbb{E}[e^{-(sD_1(k)+zD_1(k+1))}]$. Note that a downtime $D_1(k)$ can be decomposed into a waiting time $W_1(k)$ and a repair time $R_1(k)$. The waiting time $W_1(k)$ is either zero when M_2 is operational at the time of breakdown of M_1, or amounts to an exponentially (ν_2) distributed residual of the repair time of M_2 otherwise.

As noted before, machines interfere with each other in the layered model through their downtimes. More specifically, we have that a lengthy repair time of M_1 may increase the waiting time of the next downtime of M_2. At the same time, a lengthy downtime for M_2 (which might be due to a long repair time) may have an increasing influence on the next waiting time of M_1. Therefore, we have that $R_1(k)$ and $W_1(k+1)$ are positively correlated. Keeping this in mind, the bivariate LST of two consecutive downtimes is as follows:

$$\mathbb{E}[e^{-sD_1(k)-zD_1(k+1)}] = \mathbb{E}[e^{-sW_1(k)}]\mathbb{E}[e^{-zR_1(k+1)}] \int_0^\infty e^{-sy}\mathbb{E}[e^{-zW_1(k+1)}|R_1(k)=y]\nu_1 e^{-\nu_1 y}\,dy. \quad (31)$$

Since $R_1(k+1)$ is exponentially (ν_1) distributed, the terms $\mathbb{E}[e^{-sW_1(k)}]$ and $\mathbb{E}[e^{-zW_1(k+1)}|R_1(k)=y]$ remain to be computed.

First, we derive $\mathbb{E}[e^{-sW_1(k)}]$. Just before M_1 breaks down, either M_2 is up and running, or M_2 is in repair. The probability of either event happening is derived by studying the embedded Discrete Time Markov Chain (DTMC) of the machine states at epochs where any machine breaks down, starts being repaired or ends a repair period. Let $\boldsymbol{X}_n = \{X_{1,n}, X_{2,n}\}$ denote the state of the machines after the n^{th} transition. We represent the state of M_i being up at time t, waiting for repair or being in repair, by $X_{i,n} = 1$, $X_{i,n} = 2$ or $X_{i,n} = 3$ respectively. Since life times and repair times of both machines are exponential, we have that $\{\boldsymbol{X}_n, n \geq 0\}$ is a DTMC on the state space $S = \{\{1,1\},\{1,3\},\{3,1\},\{2,3\},\{3,2\}\}$. It naturally follows that the non-zero transition probabilities $p_{i,j}$ from state i to state j are given by $p_{\{1,1\},\{3,1\}} = 1 - p_{\{1,1\},\{1,3\}} = \frac{\sigma_1}{\sigma_1+\sigma_2}$, $p_{\{1,3\},\{1,1\}} = 1 - p_{\{1,3\},\{2,3\}} = \frac{\nu_2}{\sigma_1+\nu_2}$, $p_{\{3,1\},\{1,1\}} = 1 - p_{\{3,1\},\{3,2\}} = \frac{\nu_1}{\nu_1+\sigma_2}$ and $p_{\{2,3\},\{3,1\}} = p_{\{3,2\},\{1,3\}} = 1$. The DTMC is irreducible and aperiodic, hence a unique limiting distribution π on S exists and can be derived. Given this distribution function, the probability of an arbitrary transition being an event where M_1 breaks down equals $\pi_{\{1,1\}}p_{\{1,1\},\{3,1\}} + \pi_{\{1,3\}}p_{\{1,3\},\{2,3\}}$. The probability z_{up} (z_{down}) of M_2 working (being in repair), given that M_1 breaks down next transition, is thus given by

$$z_{up} = \frac{\pi_{\{1,1\}}p_{\{1,1\},\{3,1\}}}{\pi_{\{1,1\}}p_{\{1,1\},\{3,1\}} + \pi_{\{1,3\}}p_{\{1,3\},\{2,3\}}} = \frac{\sigma_1\nu_1 + (\sigma_2+\nu_1)\nu_2}{(\sigma_2+\nu_1)(\sigma_1+\sigma_2+\nu_2)},$$
$$z_{down} = \frac{\pi_{\{1,3\}}p_{\{1,3\},\{2,3\}}}{\pi_{\{1,1\}}p_{\{1,1\},\{3,1\}} + \pi_{\{1,3\}}p_{\{1,3\},\{2,3\}}} = \frac{\sigma_2(\sigma_1+\sigma_2+\nu_1)}{(\sigma_2+\nu_1)(\sigma_1+\sigma_2+\nu_2)}.$$

Hence M_1 has to wait with probability z_{down}, whereas it does not with probability z_{up}. Therefore, we have that

$$\mathbb{E}[e^{-sW_1(k)}] = z_{up} + z_{down}\frac{\nu_2}{\nu_2+s} = \frac{s\sigma_1\nu_1 + s(\sigma_2+\nu_1)\nu_2 + (\sigma_2+\nu_1)\nu_2(\sigma_1+\sigma_2+\nu_2)}{(\sigma_2+\nu_1)(s+\nu_2)(\sigma_1+\sigma_2+\nu_2)}.$$

For $\mathbb{E}[e^{-zW_1(k+1)}|R_1(k)=y]$, we first conclude that at the moment M_1 is taken into repair for y time units, M_2 must be working. After these y time units, we have a probability $e^{-\nu_2 y}$ of M_2 having broken down in the mean time, whereas it is still functioning with probability $1-e^{-\nu_2 y}$. Given the former event that M_2 is still working at the end of $R_1(k)$, there is a probability q that M_2 is in repair when M_1 breaks down again, i.e. at the start of $W_1(k+1)$. Due to the memoryless property of the exponential distribution, this probability q is easily determined by the fixed point equation

$$q = \frac{\sigma_2}{\sigma_1+\sigma_2}\left(\frac{\sigma_1}{\sigma_1+\nu_2}+\frac{\nu_2}{\sigma_1+\nu_2}q\right) \Rightarrow q = \frac{\sigma_2}{\sigma_1+\sigma_2+\nu_2}.$$

This allows us to determine the probability r that M_2 is in repair at the start of $W_1(k+1)$, given that M_2 was waiting for repair at the end of $R_1(k)$:

$$r = \frac{\sigma_1}{\sigma_1+\nu_2}+\frac{\nu_2}{\sigma_1+\nu_2}q = \frac{\sigma_1+\sigma_2}{\sigma_1+\sigma_2+\nu_2}.$$

Taking these probabilities together, we have that $W_1(k+1)$ is exponentially (ν_2) distributed with probability $e^{-\lambda_2 y}q+(1-e^{-\lambda_2 y})r$ and zero with probability $e^{-\lambda_2 y}(1-q)+(1-e^{-\lambda_2 y})(1-r)$. Thus,

$$\begin{aligned}\mathbb{E}[e^{-zW_1(k+1)}|R_1(k)=y] =& \left(e^{-\lambda_2 y}q+(1-e^{-\lambda_2 y})r\right)\frac{\nu_2}{\nu_2+z}\\ &+e^{-\lambda_2 y}(1-q)+(1-e^{-\lambda_2 y})(1-r)\\ =&e^{-\lambda_2 y}\frac{\sigma_2\nu_2+(\sigma_1+\nu_2)(\nu_2+z)}{(\sigma_1+\sigma_2+\nu_2)(\nu_2+z)}\\ &+(1-e^{-\lambda_2 y})\frac{(\sigma_1+\sigma_2)\nu_2+(\nu_2+z)\nu_2}{(\sigma_1+\sigma_2+\nu_2)(\nu_2+z)}.\end{aligned}$$

One can now compute $\mathbb{E}[e^{-(sD_1(k)+zD_1(k+1))}]$ using (31). By differentation, we obtain the moments of D_1 and the autocovariance

$$\mathrm{Cov}[D_1(k),D_1(k+1)] = \frac{\sigma_1\sigma_2}{(\sigma_2+\nu_1)^2\nu_2(\sigma_1+\sigma_2+\nu_2)}. \tag{32}$$

The correlation coefficient between $D_1(k)$ and $D_1(k+1)$ is now obtained by dividing this expression by the variance of the stationary downtime D. Now that the first two moments of the stationary downtime distribution as well as the correlation coefficient are known, we are in a position to approximate the length of Q_1 in the layered model with the result on the queue length in the single server model.

Remark 4.1. To obtain the measures required to approximate the length of Q_2 (instead of Q_1), we perform calculations similar to those above, or simply renumber the queues and machines.

Remark 4.2. Note that the covariance as given in (32) and the resulting correlation coefficient both evaluate to zero when σ_1 or σ_2 is zero, or when σ_2, ν_1 or ν_2 tends to infinity. If either σ_1 or σ_2 is zero, one of the machines essentially never breaks down and there is no interference between the machines. When σ_2 tends to infinity, there is no correlation in the downtimes of M_1 either, since M_2 is practically always down. Therefore, every single downtime of M_1 will consist of a repair time of M_1 plus a residual repair time of M_2, which are both independent of anything else. When ν_1 tends to infinity, M_1 essentially does not require any repair time from the repairman and M_2 will never have to wait for the repairman to become idle. As a result, the downtimes of M_2 are independent. A waiting time for M_1 then comes down to either zero when M_2 is up, or the residual of an M_2 repair, of which the starting time is not biased by the breakdowns of M_1. This implies that downtimes of M_1 are independent as well in that case. Equivalently, when ν_2 tends to infinity, M_2 does not require any repair time from the repairman, which means that downtimes of M_2 do not influence downtimes of M_1. As a result, there is no correlation in the M_1 downtimes in this case either. Furthermore, both the covariance and the resulting correlation coefficient are increasing in σ_1 and decreasing in both ν_1 and ν_2, due to similar arguments as the above.

Remark 4.3. In case $\sigma_1=\sigma_2$ and $\nu_1=\nu_2$, the LST $\mathbb{E}[e^{-sW_1(k)}]$ can also be obtained using the arrival theorem (cf. [18]), which states that in a closed queueing network, the stationary state probabilities at instants at which customers arrive at a service unit are equal to the stationary state probabilities at arbitrary times for the network with one less customer. This implies that at time epochs M_1 breaks down, the probability distribution on the state of M_2 (either up or in repair) is equal to the steady-state distribution of the state of M_2 in a system with $\sigma_1=0$, but with σ_2 and ν_2 left unchanged. In such a system, M_2 is the only machine requiring attention of the repairman, which greatly simplifies the analysis.

4.2 Choosing the appropriate dependence functions

In order to use Theorem 3.5 as an approximation for the PGF of L_i in the layered model, we need to identify suitable expressions for the functions $\chi_i(s)$ and $g_i(s)$. These functions need to match the dependence in the downtimes of M_i as well as possible, or equivalently, the expressions for (6) and (31) need to coincide as well as possible. The quality of the choices for the functions directly influences the accuracy of the approximation, as they are the only source of error introduced. In order to obtain suitable expressions for $\chi_i(s)$ and $g_i(s)$, we perform two-moment fits commonly used in literature. To this end, the first two moments of the distributions represented by the LST's $\chi_i(s)$ and $e^{-g_i(s)}$ must be determined. We do this based on expressions for $\chi_i'(0)$, $\chi_i''(0)$, $g_i'(0)$ and $g_i''(0)$, which we obtain by combining (4) and (7) with results for the first two moments of the downtime distribution and the correlation coefficient of the consecutive downtimes. These depend on the distributions of the repair times R_1 and R_2, among others. For exponential repair time distributions, the results required were obtained in Section 4.1 by inspection of the embedded Markov chain. For general repair time distributions, the calculation of especially the correlation coefficient may be more complex. In practice, one may obtain these numbers by statistical inspection of historical data on the downtimes of the machines.

4.2.1 Obtaining derivatives of the dependence functions

To obtain values for $\chi_i'(0)$, $\chi_i''(0)$, $g_i'(0)$ and $g_i''(0)$, we solve a set of equations. In Section 4.1, we have expressed $\mathbb{E}[D_i]$, $\mathbb{E}[D_i^2]$ and $\lim_{k\to\infty}\mathbb{E}[D_i(k)D_i(k+1)]$ in terms of the parameters of the layered model. By (4) and (7), we have that these expressions are related to the functions $\chi_i(\cdot)$ and $g_i(\cdot)$ as follows:

$$\begin{aligned}\mathbb{E}[D_i] =& \frac{\chi_i'(0)}{g_i'(0)-1},\\ \mathbb{E}[D_i^2] =& \frac{\chi_i''(0)-\mathbb{E}[D_i](2\chi_i'(0)g_i'(0)+g_i''(0))}{1-g_i'(0)^2},\\ \mathbb{E}[D_i(k)D_i(k+1)] =& -\chi_i'(0)\mathbb{E}[D_i]+g_i'(0)\mathbb{E}[D_i^2].\end{aligned}$$

These three equations in four unknowns fix values for $\chi_i'(0)$ and $g_i'(0)$, but leave one degree of freedom in the determination of $\chi_i''(0)$ and $g_i''(0)$. This freedom can be used to fine-tune the model. One might assume the independent component of the downtime to be distributed according to a certain distribution. This would lead to an additional equation for $\chi_i''(0)$ in terms of $\chi_i'(0)$, which then also fixes values for $\chi_i''(0)$ and $g_i''(0)$.

4.2.2 Expressions for the dependence functions

We now determine suitable expressions for $\chi_i(\cdot)$ and $g_i(\cdot)$. For this purpose, there are many approaches possible. Below, we base the choices of $\chi_i(\cdot)$ and $g_i(\cdot)$ on two-moment approximations. To apply these two-moment approximations, we use the notion of the squared coefficient of variation (SCV). An SCV of a random variable Z is defined to be $\frac{\mathrm{Var}[Z]}{\mathbb{E}[Z]^2}=\frac{\mathbb{E}[Z^2]}{\mathbb{E}[Z]^2}-1$.

Observe that in the previous section, we already calculated the ingredients needed to obtain the first two moments of the distributions represented by the LST's $\chi_i(s)$ and $e^{-g_i(s)}$. As explained in Section 2, the function $\chi_i(s)$ is the LST of a random variable representing the independent component of the downtime, with the first two moments given by $-\chi_i'(0)$ and $\chi_i''(0)$ respectively, and consequently with an SCV of $\frac{\chi_i''(0)}{(\chi_i'(0))^2}-1$. The function $e^{-g_i(s)}$ is the LST of an infinitely divisible distribution, the distribution of the incremental component of $D(k+1)$ per unit of $D(k)$, with as first two moments $g_i'(0)$ and $(g_i'(0))^2-g_i''(0)$ respectively, and therefore an SCV of $-\frac{g_i''(0)}{(g_i'(0))^2}$.

Based on the two moments and the SCV for each of the distributions, we employ commonly used distributional two-moment fit approximations as described in [21, p. 358–360]. For e.g. an SCV smaller than one, one fits a mixture of an Erlang(k,γ) and an Erlang$(k-1,\gamma)$ distribution to the moments ($k\geq 2, \gamma>0$), whereas for an

SCV larger than one, one uses a H_2 distribution with balanced means. In the special case of an SCV of zero or one, one uses a deterministic or exponential distribution respectively. The parameters for each of these distributions are based on the first two moments which are given as an input for this procedure.

The LSTs $\chi_i(s)$ and $e^{-g_i(s)}$ of the obtained distributional approximations now readily suggest expressions for the dependence functions. For the independent component, the LST $\chi_i(s)$ of a distributional approximation using the moments $-\chi_i'(0)$ and $\chi_i''(0)$ readily provides a suitable expression for $\chi_i(s)$. For the dependent component, the LST $e^{-g_i(s)}$ of the distributional approximation using the moments $g_i'(0)$ and $(g_i'(0))^2 - g_i''(0)$ leads to an expression for $g_i(s)$. Observe that we assumed in Section 2 that $g_i(s)$ has a completely monotone derivative, so that the LST $e^{-g_i(s)}$ represents an infinitely divisible distribution. We therefore show that the distributions used by the two moment approximation approach satisfy this assumption:

- For a deterministic distribution with value r and LST e^{-rs}, we have $g_i(s) = rs$. This function obviously has a completely monotone derivative, since $\frac{d}{ds}g_i(s) = r \geq 0$ and $\frac{d^n}{ds^n}g_i(s) = 0$ for all $n \geq 2$.
- For an exponential distribution and a H_2 distribution, see [11, p. 452] on mixtures of exponential distributions.
- A mixture of an Erlang(k, γ) and an Erlang$(k-1, \gamma)$ distribution with weights $q \in [0, 1]$ and $1-q$ respectively has LST $q\Big(\frac{\gamma}{\gamma+s}\Big)^k + (1-q)\Big(\frac{\gamma}{\gamma+s}\Big)^{k-1}$. Hence, $g_i(s) = -\log\Big(q\Big(\frac{\gamma}{\gamma+s}\Big)^k + (1-q)\Big(\frac{\gamma}{\gamma+s}\Big)^{k-1}\Big)$. Moreover we have that

  $$\frac{d^n}{ds^n}g_i(s) = (-1)^{n+1}(n-1)!\Big(\frac{k}{(\gamma+s)^n} - \frac{(1-q)^n}{(\gamma+(1-q)s)^n}\Big).$$

  The second term $(n-1)!$ is positive, as well as the third term, since $(\gamma+s)^n\frac{(1-q)^n}{(\gamma+(1-q)s)^n} \leq \frac{(\gamma+(1-q)s)^n}{(\gamma+(1-q)s)^n} = 1 < 2 \leq k$. Therefore, derivatives of odd order are positive through the first term, and negative otherwise. Hence $g_i(s)$ has a completely monotone derivative.

Example. If for the independent component we find that the SCV $\frac{\chi_i''(0)}{(\chi_i'(0))^2} - 1$ equals one, then we fit an exponential distribution with rate $(-\chi_i'(0))^{-1}$. This distribution has LST $\frac{1}{1-s\chi_i'(0)}$, which is a suitable approximation for $\chi_i(s)$. Likewise, if for the incremental component we find that the SCV $-\frac{g_i''(0)}{(g_i'(0))^2}$ equals one, again an exponential distribution is fitted with rate $(g_i'(0))^{-1}$, which provides the suggestion $e^{-g_i(s)} = \frac{1}{1+sg_i'(0)}$, or equivalently $g_i(s) = \log[1 + sg_i'(0)]$.

4.3 Resulting approximation

Now that we have obtained expressions for $\chi_i(s)$ and $g_i(s)$, Theorem 3.5 directly forms an approximation for the (PGF of the) marginal queue length distribution in the layered model:

Approximation 4.1. *In the two-layered model, an approximation $L_{i,app}$ for the queue length of Q_i is given by the PGF*

$$\mathbb{E}[p^{L_{i,app}}] = p_{up}\mathbb{E}[p^M] + p_{down}\mathbb{E}[p^{M+H(D^{past})}], \tag{33}$$

where the expressions for $p_{up}, p_{down}, \mathbb{E}[p^M]$ and $\mathbb{E}[p^{M+H(D^{past})}]$ are as given in Section 3, but with λ, $\widetilde{B}(\cdot)$, σ, $\chi(\cdot)$ and $g(\cdot)$ replaced by the two-layered model counterparts λ_i, $\widetilde{B}_i(\cdot)$, σ_i, $\chi_i(\cdot)$ and $g_i(\cdot)$.

4.4 Approximations for generalisations of the layered model

Throughout the previous sections, we derived an approximation for the layered model with two machines and a single repairman. However, the approach followed can be readily extended to approximate queue lengths of

first-layer queues in an equivalent model with a larger number of queues and machines and multiple repairmen. Moreover, the approach followed in Section 4.1 for deriving the moments and the correlation coefficient of the downtimes remains valid when assuming phase-type repair time distributions. We discuss these model generalisations below. Note that in the cases below, we only apply the analysis on the single server queueing model as given in Section 3 without any modification.

Larger numbers of machines and first-layer queues When we generalise the layered model as described in Section 2 to allow for $N > 2$ machines $M_1, \ldots, M_N$ and thus N first-layer queues $Q_1, \ldots, Q_N$, we can still use Approximation 4.1 like before to approximate $L_1, \ldots, L_N$. The approach for deriving appropriate functions for $\chi_i(s)$ and $g_i(s)$, $i = 1, \ldots, N$ needed to use Approximation 4.1 remains largely the same. However, by introducing a larger number of machines, the computation of the first two moments and the correlation coefficient of downtimes in the layered model becomes increasingly cumbersome. As opposed to the case $N = 2$ as assumed in Section 4.1, the repair buffer can now contain multiple machines. Since the repair facility serves the queue in a First-Come-First-Serve (FCFS) manner, the order in which the machines are waiting for repair needs to be included in the state space of the embedded DTMC describing the states of the machines. Subsequently, considerably more conditioning is needed to compute the terms $\mathbb{E}[e^{-sW_1(k)}]$ and $\mathbb{E}[e^{-zW_1(k+1)}|R_1(k) = y]$ in (31) and ultimately the moments and the correlation coefficient of the downtimes.

Multiple repairmen In the layered model as described in Section 2, it is assumed there is only one repairman assigned to repair machines. This assumption can be relaxed to allow for $K > 1$ repairmen in the repair facility, each working on a different machine and taking the broken machines out of the repair buffer in a FCFS manner. When $K \geq N$, a broken machine will always be taken into repair immediately. As a result, machines do not compete for repair facilities anymore, and consecutive downtimes of a machine become independent. Therefore, when taking $\chi_i(s)$ such that it equals the LST of the repair time distribution of M_i and taking $g_i(0) = 0$, the *exact* PGF of L_i is given by Approximation 4.1. When $N > K$, consecutive downtimes of the machine however remain correlated. Again, the approximation as developed in this paper remains valid, but difficulties arise in deriving the appropriate functions for $\chi_i(s)$ and $g_i(s)$, $i = 1, \ldots, N$. More specifically, the computation of the moments and the correlation coefficient of the consecutive downtimes of each of the machines becomes again increasingly cumbersome. Since machines can now be repaired simultaneously, the order in which machines return to service after repair is not necessarily the same as the order in which machines break down. This introduces extra conditioning in e.g. the computation of $\mathbb{E}[e^{-zW_1(k+1)}|R_1(k) = y]$ in (31), since the machines which were already waiting for repair at the start of $W_1(k)$ may not have returned to an operational state again by the time $R_1(k)$ has passed. This evidently influences $W_1(k + 1)$.

Phase-type distributed repair times In Section 4.1, we derived an explicit expression for the correlation coefficient of consecutive downtimes of a machine, in case repair times are exponentially distributed. Note that if we would only assume the repair time distributions to be of phase type, a similar approach for studying the embedded Markov chain can be followed to obtain the numbers needed to construct the functions $\chi_i(s)$ and $g_i(s)$ in Section 4.2. The computations may become more complex, since the current repair phase of each of the machines under repair now needs to be included in the state space of the embedded DTMC. This leads to a more complicated expression for $\mathbb{E}[e^{-sW_1(k)}]$ in (31). For the computation of $\mathbb{E}[e^{-zW_1(k+1)}|R_1(k) = y]$, extra conditioning on the repair phase is needed also. Note that even under the assumption of phase-type distributed repair times, the queue length distribution of the first-layer queues can in theory be obtained in an exact fashion, by the study of an M/G/1 queue with Markov-modulated server availability. However, due to the increased size of the state space in this embedded DTMC, the exact distribution may be very cumbersome to derive for a larger number of phases.

5 Numerical Study

In this section, we assess numerically the accuracy of the approximation that is given in Approximation 4.1. We have an initial glance at the accuracy of the approximation in Section 5.1. Then, in Section 5.2, we observe the

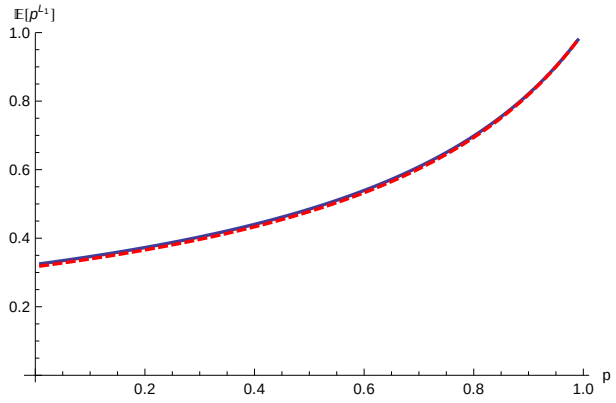

Figure 4: Plot of $\mathbb{E}[p^{L_{1,app}}]$ (solid curve) and $\mathbb{E}[p^{L_1}]$ (dashed curve).

accuracy of the approximation more thoroughly based on the numerical study of a large number of systems and we identify several key factors determining the accuracy.

5.1 Initial glance at the approximation

Before we study the accuracy of the approximation based on a large set of instances of the two-layered model as studied in this paper, we first regard the length distribution of Q_1 of a single instance of the model in order to get some initial insights. Consider a system where $\lambda_1 = 0.25$, $\sigma_1 = \sigma_2 = 1$ and B_1, R_1 and R_2 are exponentially (1) distributed. Note that the settings for λ_2 and B_2 do not influence the length of Q_1. In Figure 4 we plot the approximated PGF of $L_{1,app}$ and the 'exact' PGF of L_1 obtained by numerical methods as a function of p.

We observe in this figure that $\mathbb{E}[p^{L_{1,app}}]$ matches $\mathbb{E}[p^{L_1}]$ very closely. The error made is largest at $p = 0$, where $\mathbb{E}[p^{L_{1,app}}]$ is 2.09% larger than the value of $\mathbb{E}[p^{L_1}]$. The error decreases in p, which is why $\mathbb{E}[L]$ is approximated well. We have that $\mathbb{E}[L_{1,app}] = \frac{d}{dp}\mathbb{E}[p^{L_{1,app}}]|_{p=1} = 2.205$, while the theoretical mean $\mathbb{E}[L]$ equals 2.220.

5.2 Accuracy of the approximation

To study the accuracy of the approximation overall, we compare the approximated values for the mean of L_1 with the "exact" values (obtained by an implementation of the Power-Series Algorithm, see e.g. [4]) in various instances of the two-layered model. In particular, we regard instances where B_1 is exponentially (δ_1) distributed, and R_1 and R_2 are exponentially distributed with rates ν_1 and ν_2 respectively. The complete test bed of instances that are analysed contains 675 different combinations of parameter values, all listed in Table 1. This table lists multiple values for the workload of $Q_1(\rho_1)$, 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 as specified in the table and in the imbalance, through the values b_j^σ and b_j^ν. 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. Note that the values for ρ_2 and the service time distribution for Q_2 do not influence L_1, and hence are left unspecified.

For the systems corresponding to each of the parameter combinations in Table 1, we compare the *approximated* mean queue lengths of the first queue, $\mathbb{E}[L_{1,app}] = \frac{d}{dp}\mathbb{E}[p^{L_{1,app}}]|_{p=1}$ to the actual mean queue length $\mathbb{E}[L_1]$. Subsequently, we compute the relative error of these approximations, i.e.

$$\Delta := 100\% \times \left| \frac{\mathbb{E}[L_{1,app}] - \mathbb{E}[L_1]}{\mathbb{E}[L_1]} \right|. \tag{34}$$

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 approximation to exact results.

	0-0.01%	0.01-0.1%	0.1-1%	1-5%	>5%+
% of rel. errors	25.93%	32.30%	32.15 %	9.63 %	0.00%

Table 2: Relative errors of the mean queue length approximation applied, categorised in bins.

In Table 2 the resulting relative errors are summarised by categorisation in bins. We note that none of these errors is greater than 5%, and the majority of these errors does not exceed 1%. This shows that Approximation 4.1 approximates the mean queue length very well for the instances tested.

To observe any parameter effects, we also give the mean relative error categorised in some of the variables in Table 3. From Table 3(a), we see that the accuracy of the approximation is not very sensitive to the load of the queue. From Tables 3(b) and 3(c) we however note that the orders of magnitude of the breakdown and repair rates do impact the accuracy of the approximation. This is due to the fact that the rate at which products move (i.e., arrive and get served) with respect to the life and repair times of the machine do differ in these cases. From Tables 3(d) and 3(e), we see that the imbalance of the breakdown and repair rates do impact the accuracy as well (but to a lesser extent). We discuss the observed effects in more detail below.

Effect of fast moving products. From Table 3 we see that decreasing the uptimes and repair times of the machines relative to the movement speed of the products, or equivalently, that increasing the moving speed of products (i.e. arrival rate and service rate) relative to the uptimes and repair times, leads to a decrease in the performance of the approximation. To further examine this effect, we regard the length of Q_1 in systems with arrival rates ranging from $\lambda_1 = 0$ to $\lambda_1 = 3$ and an exponentially distributed service time B_1 with rate $\frac{10\lambda_1}{3}$ varying accordingly, so as to keep the workload fixed. Furthermore, the breakdown rates are given by $\sigma_1 = \sigma_2 = 1$ and the repair times R_1 and R_2 are exponentially (1) distributed. After applying Approximation 4.1 on the mean queue length of Q_1 in these systems and comparing it with exact results, we obtain Figure 5, where the relative error Δ (see (34)) is given as a function of λ_1. We indeed observe that the faster the products arrive (and get served), the more inaccurate the approximation becomes. This effect can be explained by the fact that faster moving products are more sensitive to variations caused by dependence in the downtimes. A small increase in the downtime, causes more additional products to build up in the queue, while such an increase may even remain unnoticed in case of slow products with long interarrival times. Hence, in the former, the error made in approximating the dependence structure of consecutive downtimes by the functions $\chi_1(\cdot)$ and $g_1(\cdot)$ shows itself more in the approximation of the mean queue length than in the latter.

Effect of the degree of dependence. From Table 3 it is apparent that the accuracy of the approximation is influenced by the values for b_j^σ and b_j^ν. This can be mainly explained by the fact that these values determine the strength of the dependence between consecutive downtimes in M_1. To illustrate this effect, let us observe systems where B_1, as well as both R_1 and R_2, is exponentially (1) distributed. Moreover, we have $\lambda_1 = 1/4$ and $\sigma_1 = 1$.

(a)

ρ_1	0.25	0.5	0.75
Mean rel. error	0.328%	0.316%	0.335%

(b)

a_i^σ	0.1	1	10
Mean rel. error	0.564%	0.294%	0.121%

(c)

a_i^ν	0.1	1	10
Mean rel. error	0.727%	0.219%	0.033%

(d)

b_j^σ	(1, 1)	(1, 2)	(2,1)	(1, 5)	(5, 1)
Mean rel. error	0.354%	0.275%	0.414%	0.149%	0.439%

(e)

b_j^ν	(1, 1)	(1, 2)	(2,1)	(1, 5)	(5, 1)
Mean rel. error	0.395%	0.344%	0.143%	0.212%	0.537%

Table 3: Mean relative error categorised in ρ_1 (a), the variables controlling the order of magnitude of σ_i and ν_i, namely a_i^σ (b) and a_i^ν (c), and the variables controlling the imbalance, b_j^σ (d) and b_j^ν (e).

In Figure 6, we show the relative error Δ in approximating the mean length of Q_1 as a function of σ_2. Since the breakdown rate of M_2 varies in these systems, we have that the strength of the dependence changes accordingly. In the figure, r_{scaled}, the correlation coefficient of consecutive downtimes as computed in Section 4.1 is given in a scaled form, so as to fit the graph. We see that the accuracy of the approximation is, at least in this case, largely determined by the strength of the correlation between the downtimes. Intuitively this makes sense, since in case there is no such correlation in the model (for example when $\sigma_2 = 0$ or $\sigma_2 \uparrow \infty$), the approximation should at least be close to being exact. Using the procedure of Section 4.2, $g_1(\cdot)$ will resolve to zero in such a case, and when $\chi_1(\cdot)$ is chosen to match $\widetilde{D}(\cdot)$, the approximation becomes exact, as the assumed downtime structure in (2) with the functions $\chi_1(\cdot)$ and $g_1(\cdot)$ will then describe the dependence in an exact way.

Comparison with Wartenhorst's approximation. The approximation approach used in the present paper involves the study of the dependence in the second layer of the model, and consequently the definition and usage of a similar, explicit dependence structure in a single server vacation queue. The two-layered model, as formulated

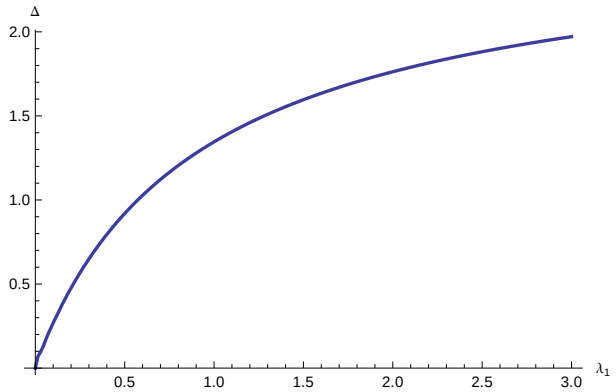

Figure 5: The relative error made as a function of the products' arrival rate.

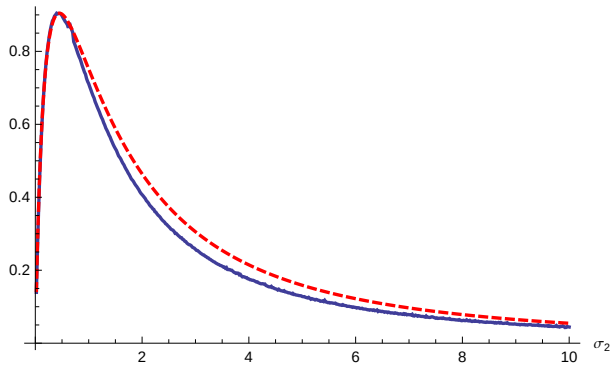

Figure 6: The relative error made, Δ (continuous curve), and the scaled value of the correlation coefficient, r_{scaled} (dashed curve) as a function of the breakdown rate of M_2.

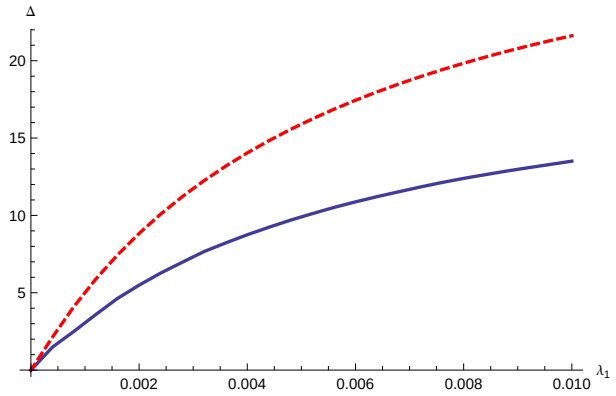

Figure 7: The relative error made by the approximation in the present paper (solid curve) and Wartenhorst's approximation (dashed curve)

in this paper, has been studied before by Wartenhorst [24]. However, it is assumed in that paper that $\sigma_1 = \sigma_2$ and that R_1 and R_2 are exponentially distributed with equal rates. This results in $\widetilde{D}_1(\cdot) = \widetilde{D}_2(\cdot)$. In his study, Wartenhorst approximates the mean length of Q_1 with the mean queue length in a single server vacation queue, where the distribution of the vacation lengths equals the stationary downtime distribution of M_i, but where the downtimes are assumed to be *completely independent*. The queue length distribution of this single server queue is obtained by applying the Fuhrmann-Cooper decomposition (cf. [13]). Wartenhorst's approximation is exact by construction for a system where downtimes are independent, and accurate whenever downtimes are only slightly dependent. Although Wartenhorst assumes equal breakdown rates and identically distributed repair times for the machines, with some effort his approach can be extended to allow for cases where these assumptions are violated. To compare the accuracy of the approximation derived in the present paper with that of Wartenhorst [24], we study a set of systems with highly dependent downtimes. For these systems, let σ_1 =100, $\sigma_2 = 0.02$, and R_2 is exponentially distributed with rate 0.01. To maximise the correlation in the downtimes of M_1, we assume R_1 to be hyperexponentially distributed with probability parameters 0.975 and 0.025, and rate parameters 100 and 0.01. The value for the correlation coefficient in these systems evaluates to 0.26. We vary λ_1 between 0 and 0.01. Furthermore, we assume B_1 to be exponentially distributed with rate $500\lambda_1$, so as to keep the workload at Q_1 fixed. In Figure 7, the relative error Δ in approximating $\mathbb{E}[L_1]$ is given for both the approximation obtained in this paper and Wartenhorst's approximation. We see the same effect of fast moving products as before. The faster the products move, the less accurate both approximations become. However, we see that the degree of dependence has a significantly larger effect on the accuracy of Wartenhorst's approximation than on that of the approximation presented in this paper. Since the degree of the dependence between the downtimes is at least the major source of inaccuracy for both approximations (cf. Section 3.4), one could conclude that the approximation derived in the

present paper performs effectively as well as Wartenhorst's approximation in cases with only slight dependences, and even better in cases with stronger correlations between the downtimes.

References

[1] E. Altman. Stochastic recursive equations with applications to queues with dependent vacations. *Annals of Operations Research*, 112:43–61, 2002.

[2] F. Baccelli and P. Brémaud. *Elements of Queueing Theory: Palm-Martingale Calculus and Stochastic Recurrences*. Springer, New York, 2003.

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

[4] 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.

[5] O.J. Boxma and M.B. Combé. The correlated M/G/1 queue. *Archiv für Elektronik und Übertragungstechnik*, 47:330–335, 1993.

[6] O.J. Boxma, M.R.H. Mandjes, and O. Kella. On a queueing model with service interruptions. *Probability in the Engineering and Informational Sciences*, 22:537–555, 2008.

[7] J.W. Cohen. *The Single Server Queue*. North-Holland, Amsterdam, 1982.

[8] B.T. Doshi. Queueing systems with vacations: a survey. *Queueing Systems*, 1:29–66, 1986.

[9] B.T. Doshi. *Single server queues with vacations. In: H. Takagi (ed.), Stochastic Analysis of Computer Communication Systems*, pages 217–265. Elsevier, Amsterdam, 1990.

[10] I. Eliazar. Gated polling systems with Lévy inflow and inter-dependent switchover times: A dynamical-systems approach. *Queueing Systems*, 49:49–72, 2005.

[11] W. Feller. *An Introduction to Probability Theory and its Applications, Vol. II*. Wiley, New York, 1971.

[12] 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.

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

[14] R. Groenevelt and E. Altman. Analysis of alternating-priority queueing models with (cross) correlated switchover times. *Queueing Systems*, 51:199–247, 2005.

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

[16] C.M. Harris and W.G. Marchal. State dependence in M/G/1 server-vacation models. *Operations Research*, 36:560–565, 1988.

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

[18] 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.

[19] R.F. Serfozo. *Introduction to Stochastic Networks*. Springer, New York, 1999.

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

[21] H.C. Tijms. *Stochastic Models: an Algorithmic Approach*. Wiley, Chichester, 1994.

[22] E.A. Van Doorn and G.J.K. Regterschot. Conditional PASTA. *Queueing Systems*, 7:229–232, 1988.

[23] W. Vervaat. On a stochastic difference equation and a representation of non-negative infinitely divisible random variables. *Advances in Applied Probability*, 11:750–783, 1979.

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

A Proof of Lemma 3.1

Proof. The function $E(p)$ is continuous on $[0, \mu(\sigma))$. We also have that $1 - E(0) = 1$ and $\lim_{p\to\mu(\sigma)} 1 - E(p) = -\infty$. Hence, there exists at least one root in $(0, \mu(\sigma))$ by Bolzano's theorem.

To prove that there is at most one root in $(0, \mu(\sigma))$, we show that $1-E(p)$ is strictly decreasing in p, or equivalently, that $E(p)$ is strictly increasing in p by studying the monotonicity of each of the terms in (17) separately. First, since $\chi(\cdot)$ is the LST of a positive continuous random variable (see Section 2), it is a strictly decreasing function. Recalling that $\lambda > 0$, this means that the first term $\chi(\lambda(1-p))$ is therefore strictly increasing in p. For the monotonicity of the second term $A^2(p)$, we show that $A(p)$ is strictly decreasing, or equivalently, $A'(p) < 0$ for all values of p considered. We have that

$$\begin{aligned} A'(p) =& \frac{\sigma\lambda}{(\sigma+\lambda(1-p))^2}\frac{p(1-\widetilde{B}(\sigma+\lambda(1-p)))}{p-\widetilde{B}(\sigma+\lambda(1-p))} \\ &+ \frac{\sigma}{\sigma+\lambda(1-p)}\Big(\frac{(1-\widetilde{B}(\sigma+\lambda(1-p)))+p\lambda\widetilde{B}'(\sigma+\lambda(1-p))}{p-\widetilde{B}(\sigma+\lambda(1-p))} \\ &\qquad - \frac{p(1-\widetilde{B}(\sigma+\lambda(1-p)))(1+\lambda\widetilde{B}'(\sigma+\lambda(1-p)))}{(p-\widetilde{B}(\sigma+\lambda(1-p)))^2}\Big). \end{aligned} \tag{35}$$

Since $\widetilde{B}(\cdot)$ is the LST of a positive, continuous random variable, we have that $1 - \widetilde{B}(\sigma + \lambda(1-p)) > 0$ and $\widetilde{B}'(\sigma+\lambda(1-p)) > 0$, which also readily implies that $p\lambda\widetilde{B}'(\sigma+\lambda(1-p)) > 0$ and $1+\lambda\widetilde{B}'(\sigma+\lambda(1-p)) > 0$. This means that in (35), the numerator of the second fraction in the first term and the numerators of the fractions between the brackets are all positive. Moreover, we have that $p - \widetilde{B}(\sigma+\lambda(1-p)) < 0$ for all $p \in (0, \mu(\sigma))$, which consequently implies through the denominators that the second fraction of the first term and the expression between the brackets each are negative. Combining this with the fact that evidently both $\sigma/(\sigma+\lambda(1-p))$ and $\sigma\lambda/(\sigma+\lambda(1-p))^2$ are positive as $p < 1$, we have that $A'(p) < 0$ and thus that the second term $A^2(p)$ is strictly increasing. For the third term $\widetilde{D}(\lambda(1-p)+g(\lambda(1-p)))$, we have that $\lambda(1-p)+g(\lambda(1-p))$ is strictly decreasing in p, for $g(s)$ is increasing in s, because $e^{-g(s)}$ is the LST of a positive continuous random variable and therefore strictly decreasing in s. Since $\lambda(1-p)+g(\lambda(1-p))$ is strictly decreasing in p and the LST $\widetilde{D}(\cdot)$ is a strictly decreasing function, the third term is strictly increasing in p.

All of the terms in (17) are strictly increasing for the values of p considered. As a result, $E(p)$ itself is strictly increasing for $p \in (0, \mu(\sigma))$. Therefore, the denominator $1 - E(p)$ has exactly one root on the real line in $(0, \mu(\sigma))$. □