Asymptotic error bounds for truncated buffer approximations of a 2-node tandem queue

Eleni Vatamidou* evatamid@gmail.com
Ivo Adan*,† i.j.b.f.adan@tue.nl
Maria Vlasiou*,‡ m.vlasiou@tue.nl
Bert Zwart*,‡ Bert.Zwart@cwi.nl

*EURANDOM and Department of Mathematics & Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
†Department of Mechanical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
‡Centrum Wiskunde & Informatica (CWI), P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

ABSTRACT

We consider the queue lengths of a tandem queueing network. The number of customers in the system can be modelled as QBD with a doubly-infinite state-space. Due to the infinite phase-space, this system does not have a product-form solution. A natural approach to find a numerical solution with the aid of matrix analytic methods is by truncating the phase-space; however, this approach imposes approximation errors. The goal of this paper is to study these approximation errors mathematically, using large deviations and extreme value theory. We obtain a simple asymptotic error bound for the approximations that depends on the truncation level. We test the accuracy of our bound numerically.

Keywords

Matrix-analytic methods; tandem queues; batch arrivals; queue length approximations; asymptotic error bound; large deviations theory; renewal theory; extreme value analysis

1. INTRODUCTION

The algorithmic evaluation of performance measures in stochastic networks is a central topic in applied probability. Indeed, many processes of interest can be modelled as Markov chains on a product space of the form $\mathbb{N} \times P$; the main coordinate of the Markov chain, called the level, is integer-valued and the phase-space P carries supplementary information. This partitioning is one of the key underlying ideas connecting phase-type distributions with algorithms that are often summarised as *Matrix-Analytic Methods* (MAM).

MAM are widely studied in the literature (see for example [7, 8, 14, 16, 24, 26, 27, 28]) and can be effective when the phase-space P is a finite set. This restriction on P limits the applicability of MAM. For example, it prevents the usage of heavy-tailed distributions as models for service times and it prevents the analysis of queueing networks with infinite waiting rooms that do not have a product form solution. Though the mathematics behind MAM can be extended to this setting using connections with the general theory of Markov additive processes [2, 25], this does not seem to lead to concrete numerical algorithms.

A natural idea to overcome this issue is simply the truncation of the phase-space P so MAM become applicable. In the examples mentioned above, this entails the approximation of heavy-tailed distributions by phase-type distributions, truncating the waiting room of a station in a queueing network, or approximating output processes by Markovian arrival processes. These ideas have in fact been applied in many engineering-oriented studies, a small sample of references being [1, 10, 13, 17, 19, 20, 30].

Somewhat surprisingly, the impact of such approximations on the accuracy of the resulting numerical algorithms is not well investigated mathematically. Classical bounds on truncation errors in Markov chains, as in [32], do not offer much insight. They are not aimed at the type of structured Markov chains encountered in queueing networks, where, for example, there is no reason to truncate the level space. The goal of this paper is to analyse mathematically the impact of truncation by means of a rigorous analysis.

Motivated by this, we consider the queue lengths of the $\mathrm{M}^X/\mathrm{M}/1 \rightarrow \bullet/\mathrm{M}/1$ tandem queueing network, where customers arrive in batches in the first queue (abbreviated as Q_1). This tandem network is a useful example of a non-product form queueing network (for non-trivial batch sizes). The number of customers in the system can be modelled as a two-dimensional Markov chain, where the marginal distribution of the number of jobs in the downstream queue (Q_2) is the hardest to obtain. For this reason, this coordinate will be chosen to be the level. A numerical solution for this model can be found by using MAM only if the buffer size of either queue is finite. For this specific model, we shall derive error bounds, with a particular emphasis on the regime where the truncation level is large, so that the resulting error is (hopefully) small.

Within the MAM literature, there have been several related works. The model we consider in this paper can be modelled as a *Quasi-Birth and Death* process (QBD) with infinite phase-space. Formally, the invariant distribution of such processes can be written as $\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 \boldsymbol{R}^i$, with $\boldsymbol{R}$ an infinite matrix [33]. A natural

MAM-9 June 28 - 30, 2016, Budapest, Hungary

DOI: 10.1145/1235

question is whether truncating the phase-space to a size N leads to a matrix $\boldsymbol{R}_N$ with the property that $\boldsymbol{R}_N \to \boldsymbol{R}$. This is also related to the question how the phase-space should be truncated: the transition probabilities of the approximating Markov chain should be augmented in such a way that the transition matrix becomes stochastic. Background on this procedure can be found in [6]. In our paper, we consider the *Partial Batch Acceptance Strategy* (PBAS), which is called *last-column augmentation* in [6]. Remarkably, this procedure does not always imply that the invariant distribution of the approximating Markov chain converges to the original one, as illustrated by Example 4.1 of [6].

Even when the invariant distribution of the approximating Markov chain converges to the original invariant distribution, one would like to know more, such as the speed of convergence. Ideally, one would like to have analytical guidelines on choosing the truncation level in such a way that a pre-described accuracy level is met. We are not aware of any analytic result in this domain. The results that seem to come closest relate to the robustness of large deviations approximations, which are in turn related to the spectral radius $\nu(N)$ of the matrix $\boldsymbol{R}_N$. There are studies showing that $\nu(N)$ does not always converge to the spectral radius ν of $\boldsymbol{R}$ [22, 31] and that the way the model is truncated actually plays a role [23].

The question examined in the present paper is closest to [6], which is to analyse the accuracy of the invariant distribution after the truncation and analyse how the error decreases when the truncation level increases. Unlike the above-mentioned works, our asymptotic techniques are based on large-deviations theory and extreme value theory, as well as Markov renewal theory. We believe that such asymptotic techniques are promising and natural to consider in this domain and have the potential to provide useful insight in the quality of numerical algorithms. This has been observed and exploited in the simulation literature (especially rare event simulation), but less so in the literature on MAM.

Specifically, our approach is as follows. Using uniformisation, we model our tandem network as a discrete time Markov chain, of which the state $(0,0)$ will be taken as regeneration point. Let $T_{(0,0)}$ be the length of a cycle and let $M^{T_{(0,0)}}$ be the maximum number of customers in the first queue during a cycle. Our first step is to show that

$$0 \leq \mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) - \mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$$
$$\leq \mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big] \frac{\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)}{\mathbb{E}T_{(0,0)}}, \qquad (1)$$

where X_∞ and Y_∞ denote the number of customers in the upstream and downstream queue in steady state, while $\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big)$, $\mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$ are the steady-state probabilities of the original and the truncated system, respectively. The first inequality is derived using a so-called Markov reward approach. The second inequality is based on arguments from regenerative process theory and essentially exploits that the original and approximating process only differ in cycles where the first queue has at least N customers. These are rather standard arguments. The main work is to analyse the asymptotic behaviour of each of the three factors on the right hand side as $N \to \infty$.

The behaviour of $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ can be reduced to studying the maximum queue length during a *small* cycle, corresponding to the busy period of the first queue in isolation. This reduction is possible using extreme value theory for regenerative processes, as surveyed in [3]. We show that we are allowed to do this by relying on ideas that date back to [18], which we adapt to the lattice case. Moreover, the term $\mathbb{E}T_{(0,0)}$ is treated in conjunction with $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$.

The behaviour of $\mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big]$ is more challenging to derive. In this paper, we give a heuristic treatment, using intuition from large deviations theory. For a formal proof we have to decompose $T_{(0,0)}$ into several (up to four) pieces, each of which we analyse using different methods. Key ingredients are optional stopping, the key Markov renewal theorem (for Markov additive processes with countable background state space) and various estimates of stopped (Markov) random walks; see e.g. [12]. Details of the proof, which is omitted for space considerations, can be found in the PhD thesis [34].

Our analysis results in a simple asymptotic estimate of the error of the form $KNe^{-\gamma N}$, where K and γ can be described explicitly in terms of the basic parameters of the model. Although our expression for γ in the leading term $e^{-\gamma N}$ seems optimal, we do not exclude that the linear term N may be removed using different arguments that are beyond the scope of this study. A numerical investigation shows that our bound may be overly conservative. Still our study seems the first to establish an asymptotic error estimate in this context.

The rest of the paper is organised as follows. In Section 2, we introduce the model under consideration and we present some additional preliminary results. In Section 3, we derive the error bounds. Furthermore, in Section 4, we derive a Cramér-Lundberg approximation for the probability $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$, which we treat together with the mean cycle length $\mathbb{E}T_{(0,0)}$. We explain intuitively in Section 5 its asymptotic behaviour. Furthermore, in Section 6, we perform numerical experiments to check the quality of the asymptotic error bound and we summarise our conclusions.

2. MODEL DESCRIPTION AND PRELIMINARIES

We consider an $\mathrm{M}^X/\mathrm{M}/1 \to \bullet/\mathrm{M}/1$ tandem queueing network. Customers arrive in batches according to a Poisson stream with rate λ and join Q_1. A customer that finishes service in Q_1 moves to Q_2. The service times for each queue are exponential with rates μ_1 and μ_2, respectively. The customer leaves the system after finishing his service in Q_2. We describe the system by a two-dimensional Markov chain $(X_n, Y_n) \in \mathbb{N}^2$, where X_n and Y_n are the queue lengths at the nth jump epoch of Q_1 and Q_2, respectively, including customers in service in either queue. For this system, we are interested in evaluating the distribution of its weak limit (X_∞, Y_∞).

We denote by B a generic r.v. of the batch sizes and we assume its mean $\mathbb{E}B = \sum_{i=1}^\infty i b_i < \infty$, where $b_i = \mathbb{P}(B = i)$, $i = 1, 2, \ldots$ Furthermore, for stability reasons, we assume that $\lambda \mathbb{E}B/\mu_i < 1$, $i = 1, 2$. In addition, w.l.o.g., we consider a uniformised version of this chain: $\lambda + \mu_1 + \mu_2 = 1$ and we denote the netput between the $(n-1)$st and the nth jump epoch in the 1st and 2nd queue as Z_n and W_n, respectively.

Namely,

$$Z_n = \begin{cases} 0, & \text{w.p.}\mu_2, \\ -1, & \text{w.p. } \mu_1, \\ m, & \text{w.p. } \lambda b_m, m = 1, 2, \dots, \end{cases} \tag{2}$$

and

$$W_n = \begin{cases} -1, & \text{if } Z_n = 0, \\ 1, & \text{if } Z_n = -1 \text{ and } X_{n-1} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Recall that due to uniformisation, $\lambda, \mu_1, \mu_2 < 1$ and the rates λ, μ_1, μ_2 can be seen as probabilities.

The number of customers X_n in Q_1 satisfies the following Lindley recursion

$$X_0 = 0, \quad X_{n+1} = \big(X_n + Z_{n+1}\big)^+, \quad n = 0, 1, \dots \tag{4}$$

Thus, $\{X_n\}_{n=0,1,\dots}$ evolves as a reflected at 0 discrete version of a random walk with increments $Z_1, Z_2, \dots$ Similarly, the number of customers Y_n in Q_2 satisfies

$$Y_0 = 0, \quad Y_{n+1} = \big(Y_n + W_{n+1}\big)^+, \quad n = 0, 1, \dots \tag{5}$$

The initial state of the system is $(X_0, Y_0) = (0, 0)$ and we define the first return time to the origin as $T_{(0,0)} = \inf\{n \geq 1 : X_n = Y_n = 0 \mid X_0 = Y_0 = 0\}$, which is also called *cycle length*. Therefore, since we have a two-dimensional positive recurrent irreducible Markov chain, it is known that

$$\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) = \frac{1}{\mathbb{E}T_{(0,0)}} \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}} \mathbb{1}\left(X_n \geq x, Y_n \geq y\right)\right].$$

From Eqs. (2) and (3), we can easily verify that the two-dimensional Markov chain (X_n, Y_n) is a QBD with an infinite phase-space $P = \{0, 1, \dots\}$, which does not admit a product form solution according to Theorem 15.1.1 of [24] unless $B = 1$.

State space truncation

As we mentioned in Section 1, the number of customers in Q_1 and Q_2 correspond to the phase and level, respectively, of the QBD introduced earlier. Thus, we truncate the buffer size of Q_1 at level N, which we call *truncation level*. More precisely, the arriving customers are admitted in the system by applying the PBAS; i.e. if the batch size is larger than the number of available free positions in the buffer (which has capacity $N - 1$), then we accept only so many customers until there are in total N customers waiting in front of Q_1 and we dismiss the remaining ones.

Moreover, we denote by $\big(X_n^{(N)}, Y_n^{(N)}\big) \in (\mathbb{N}_N \times \mathbb{N})$ the approximate Markov chain associated with the truncation level N and by $\big(Z_n^{(N)}, W_n^{(N)}\big)$ the corresponding netput process, where $\mathbb{N}_n = \{0, 1, \dots, n\} \subset \mathbb{N}$. Observe that definitions (3)–(5) are still valid (but with the notation adapted to the truncated system) for the processes $X_n^{(N)}$, $Y_n^{(N)}$, and $W_n^{(N)}$, respectively. However, the definition of $Z_{n+1}^{(N)}$ takes two alternative forms depending on the value of $X_n^{(N)}$. More precisely, if $X_n^{(N)} = N$, then

$$Z_{n+1}^{(N)} = \begin{cases} 0, & \text{w.p. } \lambda + \mu_2, \\ -1, & \text{w.p. } \mu_1, \end{cases} \tag{6}$$

while in case $X_n^{(N)} = N - m$, $m \in \{1, \dots, N\}$

$$Z_{n+1}^{(N)} = \begin{cases} 0, & \text{w.p. } \mu_2, \\ -1, & \text{w.p. } \mu_1, \\ k, & \text{w.p. } \lambda b_k \text{ for } k < m, \\ m, & \text{with probability } \lambda \sum_{i=m}^{\infty} b_i. \end{cases} \tag{7}$$

We also define $T_{(0,0)}^{(N)} = \inf\{n \geq 1 : X_n^{(N)} = Y_n^{(N)} = 0 \mid X_0^{(N)} = Y_0^{(N)} = 0\}$ as the first return time to the origin for the truncated system. Finally, we denote by $\boldsymbol{m} = (m_1, m_2)$ the two-dimensional states of the Markov chain (X_n, Y_n), where m_1 and m_2 are non-negative integers. If $\boldsymbol{P}$ is the transition probability matrix of the Markov chain and $\boldsymbol{P}^{(N)}$ its truncation, then $\forall \boldsymbol{m}, \boldsymbol{n}$ with $m_1, n_1 \in \mathbb{N}_{N-1}$ we have that

$$\boldsymbol{P}^{(N)}(\boldsymbol{m}, \boldsymbol{n}) = \boldsymbol{P}(\boldsymbol{m}, \boldsymbol{n}). \tag{8}$$

In other words, the entries in the two matrices $\boldsymbol{P}^{(N)}$ and $\boldsymbol{P}$ coincide as long as both two-dimensional Markov chains (original and truncated) live within the boundaries. This property is very useful in Section 3, where our error bounds for the approximation of the joint queue length distribution stem from this truncation.

Note that to analyse the terms $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ and $\mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big]$ (see Sections 4–5), an exponential change of measure is first required. Thus, we conclude this section by providing some results with respect to such an exponential change of measure.

Exponential change of measure

We define the *cumulant generating function* (c.g.f.) of the r.v.'s $Z_1, Z_2, \dots$ as

$$\begin{aligned} \kappa(\alpha) :&= \ln \mathbb{E}e^{\alpha Z_1} = \ln\left(\mu_2 + \mu_1 e^{-\alpha} + \lambda \mathbb{E}e^{\alpha B}\right) \\ &= \ln\left(\mu_2 + \mu_1 e^{-\alpha} + \lambda M_B(\alpha)\right), \end{aligned} \tag{9}$$

where $M_B(\alpha)$ is the *moment generating function* (m.g.f.) of the batch sizes. We assume that there exists a solution $\gamma > 0$ to the *Lundberg equation* $\kappa(\gamma) = 0$ such that $\kappa'(\gamma) < \infty$. The parameter γ is called the *adjustment coefficient* and conditions for its existence can be found in [5].

If F is the distribution of the $Z \overset{\mathfrak{D}}{=} Z_n$, we define $\breve{F}$ to be the probability distribution with density $e^{\gamma x}$ w.r.t. F, i.e. $\breve{F}(dx) = e^{\gamma x} F(dx)$ (obvious notations like $\breve{\kappa}(\alpha)$, $\breve{\mathbb{P}}$, $\breve{\mathbb{E}}$, etc, are used for quantities under the exponential change of measure). It can easily be verified that, under this exponential change of measure, the arrival rate of the batches is equal to $\breve{\lambda} = \lambda + (1 - e^{-\gamma})\mu_1$, the batch size distribution is equal to

$$\breve{\mathbb{P}}(B = n) = \frac{e^{\gamma n}}{M_B(\gamma)} \mathbb{P}(B = n), \quad n = 1, 2, \dots, \tag{10}$$

and the customers are served with rates $\breve{\mu}_1 = e^{-\gamma}\mu_1$ and $\breve{\mu}_2 = \mu_2$ in each server, respectively. In addition, it holds that $\breve{\mathbb{E}}Z = \breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1 > 0$.

We continue in the next section by providing the main results of the paper.

3. MAIN RESULTS

In this section, we present error bounds for the probability $\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big)$. In particular, we prove the two inequalities in Eq. (1). The left hand side of Eq. (1) shows that $\mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$ always

underestimates the exact probability. We formulate this result in the following proposition.

PROPOSITION 1. *If N is the truncation level of the buffer size of Q_1, then $\forall (x,y) \in \mathbb{N}^2$ it holds:*

$$\mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big) \leq \mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big). \quad (11)$$

PROOF. The proof is based on Markov reward techniques and is omitted for space considerations, for details see Section 5.3 of [34]. □

To prove the right hand side of Eq. (1), we split the steady state probability as follows

$$\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) = \frac{1}{\mathbb{E}T_{(0,0)}}\big(\mathbb{I} + \mathbb{III}\big), \quad (12)$$

$$\mathbb{I} = \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}} \mathbb{1}\left(X_n \geq x, Y_n \geq y\right) \cdot \mathbb{1}\left(\max_{1\leq l\leq T_{(0,0)}} X_l < N\right)\right],$$

$$\mathbb{III} = \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}} \mathbb{1}\left(X_n \geq x, Y_n \geq y\right) \cdot \mathbb{1}\left(\max_{1\leq l\leq T_{(0,0)}} X_l \geq N\right)\right].$$

Let $M^{T_{(0,0)}} = \max_{1\leq n\leq T_{(0,0)}} X_n$ be the maximum queue length of the first queue before the first return time to the state $(0,0)$. We show in Proposition 2 that term $\mathbb{I}$ is related to $\mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$, while term $\mathbb{III}$ evolves in some sense like $M^{T_{(0,0)}}$. With the aid of Eq. (12), we derive an upper bound for $\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big)$.

PROPOSITION 2. *An upper bound for the probability $\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big)$ is as follows:*

$$\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) \leq \mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big) + \mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big] \frac{\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)}{\mathbb{E}T_{(0,0)}}.$$

PROOF. We discuss the terms $\mathbb{I}$ and $\mathbb{III}$ separately. *Term $\mathbb{I}$:* If we set $\zeta = \inf\{n \geq 0 : X_n \geq N\}$ and $\zeta^{(N)} = \inf\{n \geq 0 : X_n^{(N)} \geq N\}$, then from Eq. (8) it holds that $(X_n : n < \zeta) \overset{\mathfrak{D}}{=} (X_n^{(N)} : n < \zeta^{(N)})$. Observe that $T_{(0,0)} = T_{(0,0)}^{(N)}$ when $\mathbb{1}\left(\max_{1\leq l\leq T_{(0,0)}} X_l < N\right) = 1$. Thus, since term $\mathbb{I}$ contains the sample paths of the truncated system, we obtain:

$$\begin{aligned}
\mathbb{I} =& \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}^{(N)}} \mathbb{1}\left(X_n^{(N)} \geq x, Y_n^{(N)} \geq y\right) \times \mathbb{1}\left(\max_{1\leq l\leq T_{(0,0)}^{(N)}} X_l^{(N)} < N\right)\right] \\
\leq& \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}^{(N)}} \mathbb{1}\left(X_n^{(N)} \geq x, Y_n^{(N)} \geq y\right)\right] \\
=& \mathbb{E}T_{(0,0)}^{(N)} \mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big). \quad (13)
\end{aligned}$$

Term $\mathbb{III}$: For the second term, we have

$$\begin{aligned}
\mathbb{III} &= \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}} \mathbb{1}\left(X_n \geq x, Y_n \geq y\right) \cdot \mathbb{1}\left(\max_{1\leq l\leq T_{(0,0)}} X_l \geq N\right)\right] \\
&\leq \mathbb{E}\left[T_{(0,0)} \cdot \mathbb{1}\left(\max_{1\leq l\leq T_{(0,0)}} X_l \geq N\right)\right] \\
&= \mathbb{E}\big[T_{(0,0)} \cdot \mathbb{1}\left(M^{T_{(0,0)}} \geq N\right)\big] \\
&= \mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big]\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big). \quad (14)
\end{aligned}$$

Combining Eqs. (12), (13), and (14), we obtain

$$\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) \leq \frac{\mathbb{E}T_{(0,0)}^{(N)}}{\mathbb{E}T_{(0,0)}}\mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big) + \frac{\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)}{\mathbb{E}T_{(0,0)}}\mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big]. \quad (15)$$

Finally, we need to show that $\mathbb{E}T_{(0,0)} \geq \mathbb{E}T_{(0,0)}^{(N)}$. Observe that $\mathbb{E}T_{(0,0)}$ and $\mathbb{E}T_{(0,0)}^{(N)}$ are by definition the expected first return times to the state $(0,0)$ in the original and the truncated system, respectively. By the strong law of large numbers for ergodic Markov chains [21], $\mathbb{E}T_{(0,0)}^{(N)} = 1/\mathbb{P}\big(X_\infty^{(N)} = 0, Y_\infty^{(N)} = 0\big)$ and $\mathbb{E}T_{(0,0)} = 1/\mathbb{P}\big(X_\infty = 0, Y_\infty = 0\big)$. Therefore, it is sufficient to show that the inequality $\mathbb{P}\big(X_\infty^{(N)} = 0, Y_\infty^{(N)} = 0\big) \geq \mathbb{P}\big(X_\infty = 0, Y_\infty = 0\big)$ holds. This inequality follows from a cost structure approach; for details see Section 5.5 of [34]. □

Observe that the term $\mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big] \times \mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)/\mathbb{E}T_{(0,0)}$ is involved in the upper bound of the steady state probability, according to Proposition 2. All factors involved in this term are hard to evaluate exactly. Instead, we examine the behaviour of these factors as $N \to \infty$.

With the aid of the exponential change of measure presented in the previous section, in Section 4, we provide asymptotic results for $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$, which is treated in conjunction with the factor $\mathbb{E}T_{(0,0)}$. Asymptotic results for the conditional expectation $\mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big]$ are derived in Section 5. The expression for the asymptotic upper bound is then formulated in Theorem 1. With $f(N) \lesssim g(N)$ we denote $\limsup_{N\to\infty} f(N)/g(N) \leq 1$.

THEOREM 1. *As $N \to \infty$,*

$$\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) - \mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big) \lesssim KNe^{-\gamma N},$$

where

$$K = \left(\frac{1}{\mu_2 - \lambda\mathbb{E}B} \cdot \left(\frac{(\breve{\mu}_1 - \mu_2)^+}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1} + \frac{(\mu_1 - \mu_2)^+}{\mu_1 - \lambda\mathbb{E}B}\right) + \frac{1}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1} + \frac{1}{\mu_1 - \lambda\mathbb{E}B}\right) \times C_1 e^\gamma \left(1 - \frac{\lambda\mathbb{E}B}{\mu_1}\right),$$

and C_1 is a constant calculated from Proposition 3.

We devote Sections 4–5 to the proof of Theorem 1.

4. ASYMPTOTIC APPROXIMATION FOR THE MAXIMUM

In this section, we derive an asymptotic approximation for $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ with the aid of extreme value theory. Observe that the number of customers in the first queue $\{X_n\}_{n=0,1,\ldots}$ forms a one-dimensional Markov chain on its own. Therefore, we denote as $T_0 = \inf\{n \geq 1 : X_n = 0 \mid X_0 = 0\}$ the return time to the origin of the first queue only and we define $M^{T_0} = \max_{1\leq n\leq T_0} X_n$. We show that the probability $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ exhibits a similar tail behaviour with

the probability $\mathbb{P}\big(M^{T_0} \geq N\big)$. Thus, we first discuss the behaviour of $\mathbb{P}\big(M^{T_0} \geq N\big)$ as $N \to \infty$.

We define $\tau_1 = \inf\{n : X_n \geq N\}$. Observe that $\mathbb{P}\big(M^{T_0} \geq N\big) = \mathbb{P}(\tau_1 < T_0)$. Moreover, the Lindley process X_n has the same transition mechanism as the random walk $U_n = Z_1 + \cdots + Z_n$, with $U_0 = 0$, until T_0, because X_n does not hit zero before T_0. Thus, it also holds that $\{\tau_1 < T_0\} = \{\tau(N-1) < \tau_-\}$, and consequently $\mathbb{P}\big(M^{T_0} \geq N\big) = \mathbb{P}(\tau(N-1) < \tau_-)$, where $\tau(N) = \inf\{n \geq 1 : U_n > N\}$ is the time of *first passage* to level $N \geq 0$ and $\tau_- = \inf\{n \geq 1 : U_n \leq 0\}$ is the first (weak) *descending ladder epoch*. We also denote the first (strict) *ascending ladder epoch* as $\tau_+ = \inf\{n \geq 1 : U_n > 0\}$. If $B(N) = U_{\tau(N)} - N$ is the *overshoot* of N, then a variant of the *Cramér-Lundberg approximation* is already known for the probability $\mathbb{P}\big(M^{T_0} \geq N\big)$ by Corollary XIII.5.9 in [4]. Therefore, we provide the following lemma without proof.

LEMMA 1. *If $B(N)$ converges in $\breve{\mathbb{P}}$ as $N \to \infty$, say to $B(\infty)$, then*

$$e^{\gamma(N-1)}\mathbb{P}\big(M^{T_0} \geq N\big) = \breve{\mathbb{E}}e^{-\gamma B(N-1)}\mathbb{1}\left(\tau_1 < T_0\right) \to C_1,$$

where $C_1 = \breve{\mathbb{P}}(\tau_- = \infty)C_0$ and $C_0 = \breve{\mathbb{E}}e^{-\gamma B(\infty)}$.

We continue by showing that the tail behaviour of $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ is similar to the tail behaviour of $\mathbb{P}\big(M^{T_0} \geq N\big)$. For this purpose, note that both $T_{(0,0)}$ and T_0 are regeneration cycles for the Markov chain X_n. Thus, if we denote $M_i^{T_0} \overset{\mathfrak{D}}{=} M^{T_0}$ as the maximum of X_n in the ith cycle T_0, where M^{T_0} is the generic cycle maximum, and similarly $M_i^{T_{(0,0)}} \overset{\mathfrak{D}}{=} M^{T_{(0,0)}}$ as the maximum of X_n in the ith cycle $T_{(0,0)}$, we have that [3, 18, 29]

$$\max_{i=1,\ldots,\frac{n}{\mathbb{E}T_{(0,0)}}} M_i^{T_{(0,0)}} \approx \max_{i=1,\ldots,n} X_i \approx \max_{i=1,\ldots,\frac{n}{\mathbb{E}T_0}} M_i^{T_0}. \tag{16}$$

We now make this precise. From Lemma 1, we know the tail behaviour of M^{T_0}. Therefore, we can derive asymptotics for the maximum $\max_{i=1,\ldots,n} X_i$. As such, Eq. (16) indicates that in order to study the asymptotic behaviour of $M^{T_{(0,0)}}$, we first need to study the asymptotics of $\max_{i=1,\ldots,n} X_i$, as $n \to \infty$.

Classically, extreme value theory focuses on finding constants a_n, b_n, such that

$$\frac{\max_{i=1,\ldots,n} X_i - a_n}{b_n} \overset{\mathfrak{D}}{\to} H, \tag{17}$$

where H is some non-degenerate r.v. and $\overset{\mathfrak{D}}{\to}$ denotes convergence in distribution. This is equivalent to showing that the probability $\mathbb{P}\big(\max_{i=1,\ldots,n} X_i \leq a_n x + b_n\big)$ has a limit, for any x. In our case, we prove that given the tail behaviour of M^{T_0} from Lemma 1, there exist constants a_n, b_n, such that (17) holds with H following the *Gumbel* function $\Lambda(x) = e^{-e^{-x}}$, $x \in \mathbb{R}$ [15].

The asymptotic behaviour of $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ is given in the following theorem. To establish this asymptotic result, we use Eq. (16) to first derive the asymptotics of $\max_{i=1,\ldots,n} X_i$, as $n \to \infty$, and later connect these asymptotics with $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$.

THEOREM 2. *It holds that*

$$\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big) \sim \frac{\mathbb{E}T_{(0,0)}}{\mathbb{E}T_0} C_1 e^{-\gamma(N-1)}, \quad N \to \infty,$$

where C_1 is defined in Lemma 1.

PROOF. The proof is based on the above-mentioned approach and is omitted for space limitations; see Section 5.5 of [34] for details. □

Observe that only the constants C_0 and C_1 are missing in order to find a closed-form asymptotic relation for the fraction $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)/\mathbb{E}T_{(0,0)}$ that appears in Eq. (1). We can find explicit expressions for these constants by using properties of lattice random walks. Thus, we conclude this section by providing explicit expressions for C_0 and C_1. We also calculate $\mathbb{E}T_0$.

Observe that both C_0 and C_1 require the evaluation of the limiting distribution of the overshoot $B(\infty)$, which can be found through the *ladder height distribution* with respect to the probability measure $\breve{\mathbb{P}}$.

Let now $\breve{H}_+$ be the distribution function of the ascending ladder height with respect to $\breve{\mathbb{P}}$ and $\breve{l}_+$ be its corresponding mean. In addition, we denote by $\breve{H}_-$ the (weak) descending ladder height distribution with respect to $\breve{\mathbb{P}}$. We have the following result.

LEMMA 2. *For a discrete-time lattice random walk, $B(\infty)$ exists with respect to $\breve{\mathbb{P}}$. In this case, C_0 is given in terms of the ladder height distributions by*

$$C_0 = \breve{\mathbb{E}}e^{-\gamma B(\infty)} = \frac{\big(1 - \|H_+\|\big)\big(1 - \|\breve{H}_-\|\big)}{(e^\gamma - 1)\kappa'(\gamma)},$$

where $\|H_+\| = \mathbb{P}(\tau_+ < \infty)$ and $\|\breve{H}_-\| = \breve{\mathbb{P}}(\tau_- < \infty)$.

PROOF. To prove this lemma, we need the limiting distribution of the overshoot, which can be obtained by adapting the renewal theorem to the lattice case, and we use Wald's equation; see Section 5.5 of [34] for details. □

PROPOSITION 3. *For a downward skip-free (or left-continuous) random walk, the constant C_1 in Lemma 1 is equal to*

$$C_1 = -\frac{\mathbb{E}Z}{\breve{\mathbb{E}}Z}(1 - e^{-\gamma})e^{-\gamma}\mu_1 = -\frac{\kappa'(0)}{\kappa'(\gamma)}(1 - e^{-\gamma})e^{-\gamma}\mu_1.$$

PROOF. From Lemma 2, it is evident that we need to find exact values for $1 - \|H_+\|$ and $1 - \|\breve{H}_-\|$. Observe that U_n is downward skip-free random walk.

We start with the evaluation of $1 - \|H_+\|$. We set $f_n = \mathbb{P}(Z = n)$. Under the probability measure $\mathbb{P}$, it holds that $\mathbb{E}Z = \kappa'(0) < 0$. Therefore, according to Corollary VIII.5.6 [4], $\|H_+\| = 1 + \mathbb{E}Z/f_{-1}$, where from Eq. (3) we know that $f_{-1} = \mathbb{P}(Z = -1) = \mu_1$.

By the definition of the descending ladder height distribution, we have that

$$1 - \|\breve{H}_-\| = \breve{\mathbb{P}}(\tau_- = \infty) = \breve{\mathbb{P}}(U_n \geq 1 \text{ for all } n \geq 1).$$

We set now $\breve{f}_n = \breve{\mathbb{P}}(Z = n)$ and $T_1 = \inf\{n : U_n = -1\}$. Since U_n is a downward skip-free random walk with an upward drift under the probability measure $\breve{\mathbb{P}}$, it holds from Proposition 11 in [9] that

$$1 - \|\breve{H}_-\| = \breve{f}_{-1} \cdot \frac{1 - \breve{\mathbb{P}}(T_1 < \infty)}{\breve{\mathbb{P}}(T_1 < \infty)}.$$

Thus, it is left to find the probability $\breve{\mathbb{P}}(T_1 < \infty)$, which according to Lemma 2 in [9] is equal to the unique value $s \in (0,1)$ that satisfies the equation $\breve{\mathbb{E}}s^Z = 1$. Using $\kappa(\alpha) = 0$, we get from Proposition XIII.1.1 in [4] that $\breve{\mathbb{E}}e^{\alpha Z_1} = e^{\kappa(\alpha+\gamma)}$. Therefore, $\breve{\mathbb{E}}e^{-\gamma Z} = e^{\kappa(0)} = 1$, and consequently $s = e^{-\gamma} \in (0,1)$ is the unique solution to

the equation $\breve{\mathbb{E}}s^Z = 1$. As a result, $\breve{\mathbb{P}}(T_1 < \infty) = e^{-\gamma}$. We also find $\breve{f}_{-1} = \breve{\mathbb{P}}(Z = -1) = e^{-\gamma}\mu_1$. Combining all the above and Lemma 1, the result is immediate. □

We turn now our attention to the evaluation of $\mathbb{E}T_0$. Observe that $\mathbb{E}T_0 = 1/\mathbb{P}\big(X_\infty = 0\big)$. By applying Little's law for a busy server we find that $\rho_1 = \lambda\mathbb{E}B/\mu_1$, with $\lambda\mathbb{E}B$ being the average number of customers entering the system per time unit. Consequently, $\mathbb{P}\big(X_\infty = 0\big) = 1 - \rho_1 = 1 - \lambda\mathbb{E}B/\mu_1$. Thus, we have proven:

LEMMA 3. $\mathbb{E}T_0 = \big(1 - \lambda\mathbb{E}B/\mu_1\big)^{-1}$.

5. THE CONDITIONAL MEAN RETURN TIME

Our last goal is to study the asymptotic behaviour of the conditional expectation $\mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big]$. More precisely, we study the limit $\lim_{N\to\infty} \frac{1}{N}\mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big]$. We take a heuristic approach, using intuition from large deviations theory. A formal proof can be found in Section 5.6 of [34]. Define

- τ_1: the time at which Q_1 reaches or exceeds level N. Recall that it was defined in Section 4 as $\tau_1 = \inf\{n : X_n \geq N\}$.
- τ_2: the return time to 0 in Q_1 after τ_1. Formally, $\tau_2 = \inf\{n > \tau_1 : X_n = 0\}$.
- τ_3: the first time Q_2 empties after τ_2. Formally, $\tau_3 = \inf\{n > \tau_2 : Y_n = 0\}$. The time τ_3 can either coincide with or happen before $T_{(0,0)}$.

We describe heuristically how both queues behave, given that the number of customers in Q_1 has reached a very high level before the first return time $T_{(0,0)}$ to the empty state $(0,0)$. Our description is based on intuition from large deviations theory and fluid limits. We write $a \approx b$ to denote that a is approximately equal to b, without explicitly determining the degree of accuracy. Denote by $\#Q_1$ and $\#Q_2$ the number of customers in Q_1 and Q_2.

Observe that the behaviour of Q_1 is not affected by what happens in Q_2. On the other hand, we recognise three different cases for the behaviour of Q_2 that arise from the relation between the rates μ_1, μ_2, and $\breve{\mu}_1$. We summarise all cases in Figure 1. We start by discussing the behaviour of Q_1.

To describe the behaviour of Q_1 until time $T_{(0,0)}$, given that $\#Q_1$ reached or exceeded level N, we apply arguments from large deviations theory. According to Section 2, this event happens by a change of measure, from $\mathbb{P}$ to $\breve{\mathbb{P}}$. Since $N \to \infty$, the time its takes Q_1 from τ_1 to reach its maximum value (something above N) before $T_{(0,0)}$ is negligible (compared to τ_1). Moreover, until τ_1, the departure rate of the customers is asymptotically equal to $\breve{\mu}_1$ because the system is overloaded ($\breve{\lambda}\breve{\mathbb{E}}B > \breve{\mu}_1$). On the other hand, after τ_1, all the rates are back to normal. As we have already mentioned, τ_2 is the point at which the Q_1 reaches 0 after reaching its maximum value within cycle $T_{(0,0)}$. Since during the time interval $[\tau_1, \tau_2]$ Q_1 is always full, the departure rate of customers equals μ_1.

Next, we describe the behaviour of Q_2 before $T_{(0,0)}$.

Case 1: $\mu_1 < \mu_2$

It always holds that $\breve{\mu}_1 < \mu_1$ (see Section 2 for the definition of $\breve{\mu}_1$). Therefore, in this case, Q_2 behaves asymptotically as a stable M/M/1 queue in all time intervals (but with different arrival rates of customers). Thus, the number of customers in Q_2 is bounded by the number of customers in a stable M/M/1 queue until $T_{(0,0)}$. Consequently, the time interval $[\tau_2, T_{(0,0)}]$ is negligible compared to $[0, \tau_2]$ and we expect that $\mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big] \approx \mathbb{E}\big[\tau_2 \mid M^{T_{(0,0)}} \geq N\big]$, where from Euclidean geometry we can easily verify that (see Figure 1)

$$\tau_1 \approx \frac{N}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1}, \qquad \tau_2 - \tau_1 \approx \frac{N}{\mu_1 - \lambda\mathbb{E}B}. \tag{18}$$

Case 2: $\breve{\mu}_1 < \mu_2 < \mu_1$

Since $\breve{\mu}_1 < \mu_2$, Q_2 behaves asymptotically as a stable M/M/1 queue with arrival rate $\breve{\mu}_1$ and service rate μ_2 until time τ_1. This means that the number of customers in Q_2 at time τ_1 is bounded by the number of customers in the latter M/M/1 queue. From τ_1 onwards, the arrival rate of customers in Q_2 is equal to μ_1, which is greater than the service rate μ_2. Therefore, the number of customers in Q_2 grows linearly with rate $\mu_1 - \mu_2$ up until τ_2. After τ_2, the output rate from Q_1 is equal to $\lambda\mathbb{E}B$ and the customers in Q_2 reduce linearly with rate $\lambda\mathbb{E}B - \mu_2$ until the queue empties at time τ_3. We calculate (see Figure 1)

$$\begin{aligned} h_2 &\approx (\mu_1 - \mu_2)\frac{N}{\mu_1 - \lambda\mathbb{E}B}, \\ \tau_3 - \tau_2 &\approx \frac{h_2}{\mu_2 - \lambda\mathbb{E}B} = \frac{\mu_1 - \mu_2}{\mu_2 - \lambda\mathbb{E}B} \cdot \frac{N}{\mu_1 - \lambda\mathbb{E}B}. \end{aligned} \tag{19}$$

Obviously, in this case $\mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big] \approx \mathbb{E}\big[\tau_3 \mid M^{T_{(0,0)}} \geq N\big]$, because the interval $[\tau_3, T_{(0,0)}]$ is negligible compared to $[0, \tau_3]$.

Case 3: $\mu_2 < \breve{\mu}_1 < \mu_1$

Since $\breve{\mu}_1 > \mu_2$, the number of customers in Q_2 grows linearly with rate $\breve{\mu}_1 - \mu_2$ up until time τ_1. For the remaining time intervals, Q_2 behaves in a similar manner as in Case 2. Therefore, $\mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big] \approx \mathbb{E}\big[\tau_3 \mid M^{T_{(0,0)}} \geq N\big]$, where (see Figure 1)

$$\begin{aligned} h_1 &\approx (\breve{\mu}_1 - \mu_2)\frac{N}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1}, \\ h_2 &\approx h_1 + (\mu_1 - \mu_2)\frac{N}{\mu_1 - \lambda\mathbb{E}B}, \\ \tau_3 - \tau_2 &\approx \frac{h_2}{\mu_2 - \lambda\mathbb{E}B}. \end{aligned} \tag{20}$$

To prove rigorously the behaviour of Q_2 in $[0, \tau_2]$, we use renewal theory arguments and the relation between $\mathbb{P}$ and $\breve{\mathbb{P}}$. For the time interval $[\tau_2, \tau_3]$, the idea is to see our two-dimensional Markov chain as a *Markov Additive Process* (MAP) [11]. Finally, for $[\tau_3, T_{(0,0)}]$, we use that the hitting time of the origin is finite since the latter is a recurrent state for our ergodic Markov chain.

6. NUMERICAL EXPERIMENTS

We perform now numerical experiments to check the quality of our asymptotic upper error bound (*a.u.e.b.*) in Theorem 1. As an example, we use geometric batch sizes, where we calculate the exact queue lengths through simulation.

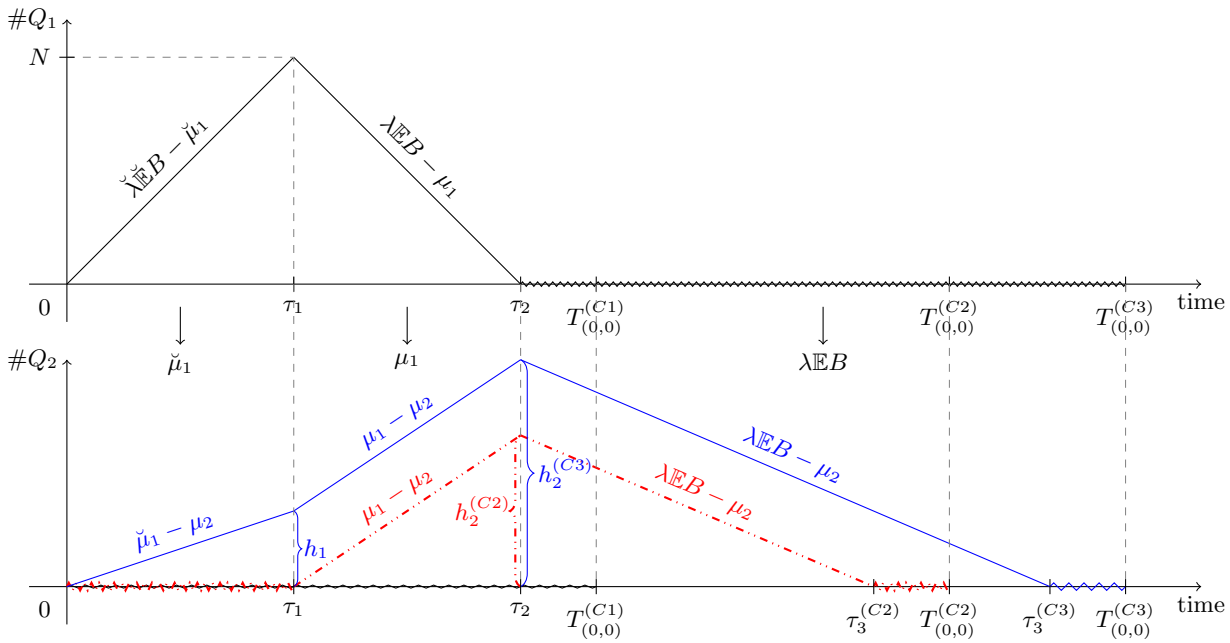

Figure 1: The asymptotic behaviours of Q_1 and Q_2, given that $\#Q_1$ before $T_{(0,0)}$ exceeded the truncation level N, for all 3 different cases; solid black for Case 1, dash-dotted red for Case 2, and solid blue for Case 3.

Suppose that $\mathbb{P}(B=n)=\beta(1-\beta)^{n-1}$, $n=1,2,\dots$ We find $\gamma=-\ln\big((\lambda+\mu_1-\beta\mu_1)/\mu_1\big)$ and the rates with respect to the measure $\breve{\mathbb{P}}$ take the form $\breve{\lambda}=\beta\mu_1$ and $\breve{\mu}_1=\lambda+\mu_1-\beta\mu_1$. We also find that $\breve{\mathbb{E}}B=(\lambda+\mu_1-\beta\mu_1)/\lambda$. Finally, using Proposition 3, we also calculate that $C_1=(\beta\mu_1-\lambda)\lambda/\beta\mu_1$. Combining these expressions, we calculate the *a.u.e.b.* in Theorem 1.

For our numerical experiments, we focus on the marginal distribution of Q_2. We performed extensive numerical experiments for various combinations of the parameters. We present here the combinations $\{\beta=0.5,\rho_1=0.7,\rho_2=0.8\}$ (Case 2), since qualitatively they represent the most pessimistic case among the various combinations we tested. Observe that due to the uniformisation $\lambda+\mu_1+\mu_2=1$ of the rates, there exists a unique combination of $\{\lambda,\mu_1,\mu_2\}$ given a combination $\{\beta,\rho_1,\rho_2\}$. For this combination, we choose a number of truncation levels and we calculate for each N the truncated approximation $\mathbb{P}\big(Y_\infty^{(N)}\geq y\big)$, $y\geq 0$, with MAM.

To check the quality of our *a.u.e.b.*, we compare it with the differences between the exact and the truncated approximation of the marginal distribution of Q_2. We summarise our findings in Table 1.

From the table, we observe that the truncated approximations become more accurate as N increases, which is in accordance with our expectations. The same also holds for the asymptotic bound. However, the bound is at least 5 times greater than the observed error, which makes it rather conservative.

Similar results were derived in [34], where we performed additional numerical experiments for the special case of single arrivals of customers.

y	$N=10$	$N=20$	$N=30$	$N=50$
5	0.128921	0.025536	0.005539	0.000755
10	0.123171	0.029763	0.006556	0.000517
15	0.086761	0.026535	0.006317	0.000349
20	0.054454	0.020534	0.005432	0.000237
25	0.032516	0.014616	0.004358	0.000221
30	0.018948	0.009835	0.003276	0.000195
a.u.e.b.	0.617191	0.243018	0.018839	0.004636

Table 1: Observed errors between the original marginal distribution of Q_2 and its QBD approximation for $\rho_1=0.7$ and $\rho_2=0.8$. The last line corresponds to the *a.u.e.b* for each N.

7. CONCLUSIONS

The conclusions we can draw for the asymptotic upper bound are summarised as follows: (i) The bound depends only on the truncation level and the parameters of the model; i.e. it is uniform in the values x and y of $\mathbb{P}\big(X_\infty^{(N)}\geq x,Y_\infty^{(N)}\geq y\big)$. (ii) The bound is rather conservative. Moreover, the bound becomes more conservative as the truncation level increases. (iii) Given that it seems impossible to improve upon the leading factor $e^{-\gamma N}$, the conservative behaviour that our bound exhibits is probably attributed to the factor N.

The above observations indicate that further modifications are important to improve the accuracy of the asymptotic upper bound. One possible direction is to make the bound dependent on the values x and y. Most importantly, since the factor N of the bound seems to be more responsible for the latter's conservative behaviour, further improvements should be sought towards the removal of this factor from the bound. Nonetheless, the advantage of our bound is clear, in that it makes the procedure of truncating the background state rigorous while leading to an asymptotic expression that converges to zero.

8. ACKNOWLEDGMENTS

The work of Maria Vlasiou and Eleni Vatamidou is supported by Netherlands Organisation for Scientific Research (NWO) through project number 613.001.006. The work of Bert Zwart is supported by an NWO VICI grant.

9. REFERENCES

[1] A. Alfa, B. Liu, and Q. He. Discrete-time analysis of $MAP/PH/1$ multiclass general preemptive priority queue. *Naval Research Logistics*, 50(6):662–682, 2003.

[2] E. Arjas and T. P. Speed. Symmetric Wiener-Hopf factorisations in Markov additive processes. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 26:105–118, 1973.

[3] S. Asmussen. Extreme value theory for queues via cycle maxima. *Extremes. Statistical Theory and Applications in Science, Engineering and Economics*, 1(2):137–168, 1998.

[4] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, New York, 2003.

[5] S. Asmussen and H. Albrecher. *Ruin Probabilities*. Advanced Series on Statistical Science & Applied Probability, 14. World Scientific, Second edition, 2010.

[6] N. Bean and G. Latouche. Approximations to quasi-birth-and-death processes with infinite blocks. *Advances in Applied Probability*, pages 1102–1125, 2010.

[7] D. A. Bini, G. Latouche, and B. Meini. *Numerical Methods for Structured Markov Chains*. Numerical Mathematics and Scientific Computation. Oxford University Press, 2005.

[8] L. Breuer and D. Baum. *An Introduction to Queueing Theory: and Matrix-Analytic Methods*. Springer, 2005.

[9] M. Brown, E. A. Peköz, and S. M. Ross. Some results for skip-free random walk. *Probability in the Engineering and Informational Sciences*, 24(4):491–507, 2010.

[10] G. Casale, E. Z. Zhang, and E. Smirni. KPC-Toolbox: Best recipes for automatic trace fitting using Markovian Arrival Processes. *Performance Evaluation*, 67(9):873–896, 2010.

[11] E. Çinlar. Markov additive processes. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 24(2):85–93, 1972.

[12] R. Doney, R. Maller, and M. Savov. Renewal theorems and stability for the reflected process. *Stochastic Processes and their Applications*, 119(4):1270 – 1297, 2009.

[13] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31(3–4):245–279, 1998.

[14] H. R. Gail, S. L. Hantler, and B. A. Taylor. Spectral analysis of $M/G/1$ and $G/M/1$ type Markov chains. *Advances in Applied Probability*, 28(1):114–165, 1996.

[15] E. J. Gümbel. *Statistics of extremes*. Columbia University Press, 1958.

[16] Q. M. He. *Fundamentals of Matrix-Analytic Methods*. Springer, 2014.

[17] A. Horváth, G. Horváth, and M. Telek. A joint moments based analysis of networks of MAP/MAP/1 queues. *Performance Evaluation*, 67(9):759–778, 2010.

[18] D. L. Iglehart. Extreme values in the GI/G/1 queue. *The Annals of Mathematical Statistics*, 43(2):627–635, 1972.

[19] E. Kao and K. Narayanan. Modeling a multiprocessor system with preemptive priorities. *Management Science*, 37(2):185–197, 1991.

[20] A. Kapadia, M. Kazmi, and A. Mitchell. Analysis of a finite capacity non preemptive priority queue. *Computers & Operations Research*, 11(3):337–343, 1984.

[21] M. Kijima. *Markov processes for stochastic modeling*. Springer, 1997.

[22] D. P. Kroese, W. R. W. Scheinhardt, and P. G. Taylor. Spectral properties of the tandem Jackson network, seen as a Quasi-Birth-and-Death process. *The Annals of Applied Probability*, 14(4):2057–2089, 2004.

[23] G. Latouche, G. Nguyen, and P. Taylor. Queues with boundary assistance: the effects of truncation. *Queueing Systems*, 69(2):175–197, 2011.

[24] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. SIAM, 1999.

[25] M. Miyazawa and B. Zwart. Wiener-Hopf factorizations for a multidimensional Markov additive process and their applications to reflected processes. *Stochastic Systems*, 2(1):67–114, 2012.

[26] M. F. Neuts. *Structured Stochastic Matrices of $M/G/1$ Type and their Applications*, volume 5 of *Probability: Pure and Applied*. Marcel Dekker Inc., 1989.

[27] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models*. Dover Publications Inc., 1994. Corrected reprint of the 1981 original.

[28] A. Ost. *Performance of communication systems: a model-based approach with matrix-geometric methods*. Springer, 2001.

[29] H. Rootzén. Maxima and exceedances of stationary Markov chains. *Advances in Applied Probability*, 20(2):371–390, 1988.

[30] R. Sadre. *Decomposition-Based Analysis of Queueing Networks*. PhD thesis, University of Twente, 2007.

[31] Y. Sakuma and M. Miyazawa. On the effect of finite buffer truncation in a two-node Jackson network. *Journal of Applied Probability*, 42(1):199–222, 2005.

[32] E. Seneta. *Nonnegative Matrices and Markov Chains*. Springer Series in Statistics. Springer-Verlag, Second edition, 1981.

[33] R. L. Tweedie. Operator-geometric stationary distributions for Markov chains, with application to queueing models. *Advances in Applied Probability*, 14(2):368–391, 1982.

[34] E. Vatamidou. *Error analysis of stuctured Markov chains*. PhD thesis, Eindhoven University of Tehcnology, 2015.