Accepted Manuscript

Occupation times for the finite buffer fluid queue with phase-type ON-times

N.J. Starreveld, R. Bekker, M. Mandjes

PII: S0167-6377(16)30289-9
DOI: https://doi.org/10.1016/j.orl.2017.10.012
Reference: OPERES 6290

To appear in: *Operations Research Letters*

Received date : 23 December 2016
Revised date : 19 October 2017
Accepted date : 19 October 2017

Please cite this article as: N.J. Starreveld, R. Bekker, M. Mandjes, Occupation times for the finite buffer fluid queue with phase-type ON-times, *Operations Research Letters* (2017), https://doi.org/10.1016/j.orl.2017.10.012

Occupation times for the finite buffer fluid queue with phase-type ON-times

N. J. Starreveld, R. Bekker and M. Mandjes

Abstract

In this short communication we study a fluid queue with a finite buffer. The performance measure we are interested in is the occupation time over a finite time period, i.e., the fraction of time the workload process is below some fixed target level. Using an alternating renewal sequence, we determine the double transform of the occupation time; the occupation time for the finite buffer M/G/1 queue with phase-type jumps follows as a limiting case.

Keywords: Occupation time ∘ fluid model ∘ phase-type distribution ∘ doubly reflected process ∘ finite buffer queue

Affiliations:. N. J. Starreveld is with Korteweg-de Vries Institute for Mathematics, Science Park 904, 1098 XH Amsterdam, University of Amsterdam, the Netherlands. Email: n.j.starreveld@uva.nl. R. Bekker is with Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands. Email: r.bekker@vu.nl. M. Mandjes is with Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands; he is also affiliated with EURANDOM, Eindhoven University of Technology, Eindhoven, the Netherlands, and CWI, Amsterdam the Netherlands. Email: m.r.h.mandjes@uva.nl.

1. Introduction

Owing to their tractability, the OR literature predominantly focuses on queueing systems with an *infinite* buffer or storage capacity. In practical applications, however, we typically encounter systems with *finite*-buffer queues. Often, the infinite-buffer queue is used to approximate its finite-buffer counterpart, but it is questionable whether this is justified when the buffer is not so large.

In this paper we consider the workload process $\{Q(t)\}_{t\geq 0}$ of a fluid queue with finite workload capacity $K>0$. Using the results for the fluid queue we also analyze the finite-buffer M/G/1 queue. The performance measure we are interested in is the so-called *occupation time* of the set $[0,\tau]$ up to time t, for some $\tau\in[0,K]$, defined by

$$\alpha(t):=\int_0^t 1_{\{Q(s)\in[0,\tau]\}}\mathrm{d}s. \tag{1.1}$$

Our interest in the occupation time can be motivated as follows. The queueing literature mostly focuses on stationary performance measures (e.g. the distribution of the workload $Q(t)$ when $t\to\infty$) or on the performance after a finite time (e.g. the distribution of $Q(t)$ at a fixed time $t\geq 0$). Such metrics do not always provide operators with the right means to assess the service level agreed upon with their clients. Consider for instance a call center in which the service level is measured over intervals of several hours during the day; a typical service-level target is then that 80% of the calls should be answered within 20 seconds. Numerical

results for this call center setting [5, 14, 15] show that there is severe fluctuation in the service level, even when measured over periods of several hours up to a day. The numerical results in Section 4 indicate that this also holds for single-server queues with finite buffer. Using a stationary measure for the average performance over a finite period may thus be highly inadequate (unless the period over which is averaged is long enough). Our work is among the first attempts to study occupation times in finite-capacity queueing systems.

Whereas there is little literature on occupation times for queues, there is a substantial body of work on occupation times in a broader setting. One stream of research focuses on occupation times for processes whose paths can be decomposed into regenerative cycles [7, 16, 18, 19]. Another branch is concerned with occupation times of spectrally negative Lévy processes, see e.g. [12, 13]. The results established typically concern occupation times until a first passage time, whereas [11] focuses on refracted Lévy processes. In [16] spectrally positive Lévy processes with reflection at the infimum were studied as a special case; we also refer to [16] and references therein for additional literature. To the best of our knowledge there is no paper on occupation times for doubly reflected processes, as we consider here.

In this paper we use the framework studied in [16]. More specifically, the occupation time is cast in terms of an alternating renewal process, whereas for the current setting the upper reflecting barrier complicates the analysis. We consider a finite buffer fluid queue where during ON times the process increases linearly and during OFF times the process decreases linearly. We consider the case that ON times have a phase-type distribution and the OFF times have an exponential distribution. This framework allows us to exploit the regenerative structure of the workload process and provides the finite-capacity M/G/1 queue with phase-type jumps as a limiting special case. For this model we succeed in deriving closed-form results for the Laplace transform (with respect to t) of the occupation time. Relying on the ideas developed in [3], all quantities of interest can be explicitly computed as solutions of systems of linear equations.

The structure of the paper is as follows. In Section 2 we describe the model and give some preliminaries. Our results are presented in Section 3. A numerical implementation of our method and some numerical experiments are presented in Section 4. A more detailed version of this short communication can be found on arXiv, [17].

2. Model description and preliminaries

We consider the finite capacity fluid queue with linear rates. The rate is determined by an independently evolving Markov chain, where we assume that there is only one state in which work decreases; this may be interpreted as the OFF time of a source that feeds work into the queue. There are multiple states of the underlying Markov chain during which work accumulates at (possibly) different linear rates. In case these rates are identical, periods during which work accumulates may be interpreted as ON times of a corresponding ON-OFF source. The ON-OFF source then has exponentially distributed OFF times, whereas the ON times follow a phase-type distribution. The workload capacity is K and work that does not fit is lost; see Subsection 2.2 for a more formal description. Some basic results concerning phase-type distributions and martingales that are used in the sequel are first presented in Subsection 2.1.

2.1. Preliminaries

Phase-type distributions. A phase-type distribution B is defined as the *absorption time* of a continuous-time Markov process $\{\mathcal{J}(t)\}_{t\geq 0}$ with finite state space $E \cup \{\partial\}$ such that ∂ is an absorbing state and the states in E are transient. We denote by $\boldsymbol{\alpha}_0$ the initial probability distribution of the Markov process, by $\mathbf{T}$ the *phase generator* and by $\mathbf{t}$ the *exit vector*. The vector $\mathbf{t}$ can be equivalently written as $-\mathbf{T}\mathbf{1}$, where $\mathbf{1}$ is a column vector

of ones. We denote such a phase-type distribution by $(n, \boldsymbol{\alpha}_0, \mathbf{T})$ where $|E| = n$. For simplicity we assume that $E = \{1, \ldots, n\}$. In what follows we denote by B a phase-type distribution with representation $(n, \boldsymbol{\alpha}_0, \mathbf{T})$; for a phase-type distribution with representation $(n, \mathbf{e}_i, \mathbf{T})$ we add the subscript i in the notation. An important property of the class of phase-type distributions is that it is dense (in the sense of weak convergence) in the set of all probability distributions on $(0, \infty)$; see [2, Thm. 4.2]. For a phase-type distribution with representation $(n, \boldsymbol{\alpha}_0, \mathbf{T})$, the cumulative distribution function $B(\cdot)$, the density $b(\cdot)$ and the Laplace transform $\hat{B}(\cdot)$ are given in [2, Prop. 4.1]. In particular, for $x \geq 0$ and $s \geq 0$, we have

$$\mathbb{P}(B > x) = \boldsymbol{\alpha}_0^{\mathrm{T}} e^{\mathbf{T}x} \mathbf{1} \quad \text{and} \quad \hat{B}(s) = \boldsymbol{\alpha}_0^{\mathrm{T}} (s\mathbf{I} - \mathbf{T})^{-1} \mathbf{t}. \tag{2.1}$$

When the phase-type distribution has representation $(n, \mathbf{e}_i, \mathbf{T})$ we use the notation $\hat{B}_i(\cdot)$ instead of $\hat{B}(\cdot)$. For a general overview of the theory of phase-type distributions we refer to [2, 3] and references therein.

Markov-additive fluid process (MAFP). Markov-additive fluid processes belong to a more general class of processes called *Markov-additive processes*, see [2, Ch. XI]. Consider a right-continuous irreducible Markov process $\{\mathcal{J}(t)\}_{t\geq 0}$ defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a finite state space $E = \{1, \ldots, n\}$ and rate transition matrix $\mathcal{Q}$. While the Markov process $\mathcal{J}(\cdot)$ is in state i the process $X(\cdot)$ behaves like a linear drift r_i. We assume that the rates $r_1, \ldots, r_n$ are independent of the process $\mathcal{J}(\cdot)$. Such a process will be referred to as a *Markov-additive fluid process* and abbreviated as MAFP. For $z \in \mathbb{C}^{\mathrm{Re}\geq 0}$, the *matrix exponent* of the MAFP is defined as

$$F(z) = \mathcal{Q} - z\,\mathrm{diag}(r_1, \ldots, r_n) = \mathcal{Q} - z\boldsymbol{\Delta}_r, \tag{2.2}$$

where $\boldsymbol{\Delta}_r = \mathrm{diag}(r_1, \ldots, r_n)$. In what follows we shall need information concerning the roots of the equation

$$\det\left(F(z) - q\mathbf{I}\right) = \det(\mathcal{Q} - z\boldsymbol{\Delta}_r - q\mathbf{I}) = 0, \tag{2.3}$$

with $q \geq 0$. From [9] we have that there exist n values $\rho_1(q), \ldots, \rho_n(q)$ and corresponding vectors $\mathbf{h}_1(q), \ldots, \mathbf{h}_n(q)$ such that, for each $k = 1, \ldots, n$, $\det\left(\mathcal{Q} - \rho_k(q)\boldsymbol{\Delta}_r - q\mathbf{I}\right) = 0$ and $(\mathcal{Q} - \rho_k(q)\boldsymbol{\Delta}_r - q\mathbf{I})\mathbf{h}_k(q) = 0$.

The Kella-Whitt martingale. The counterpart of the Kella-Whitt martingale for *Markov-additive processes* was established in [4]; let $\{Y(t)\}_{t\geq 0}$ be an adapted continuous process having finite variation on compact intervals. Set $Z(t) = X(t) + Y(t)$ and let $z \in \mathbb{C}^{\mathrm{Re}\geq 0}$. Then, for every initial distribution $(X(0), \mathcal{J}(0))$,

$$\mathbf{M}(z,t) := \int_0^t e^{-zZ(s)} \mathbf{e}_{\mathcal{J}(s)} \mathrm{d}s\; F(z) + e^{-zZ(0)} \mathbf{e}_{\mathcal{J}(0)} - e^{-zZ(t)} \mathbf{e}_{\mathcal{J}(t)} - z \int_0^t e^{-zZ(s)} \mathbf{e}_{\mathcal{J}(s)} \mathrm{d}Y(s) \tag{2.4}$$

is a vector-valued zero mean martingale.

2.2. Fluid model with two reflecting barriers

The MAFP $(X(t), \mathcal{J}(t))_{t\geq 0}$ we analyze has a modulating Markov process $\{\mathcal{J}(t)\}_{t\geq 0}$ with state space $E = \{1, \ldots, n+1\}$ and generator $\mathcal{Q}$ given by

$$\mathcal{Q} = \begin{bmatrix} -\lambda & \lambda \boldsymbol{\alpha}_0^{\mathrm{T}} \\ \mathbf{t} & \mathbf{T} \end{bmatrix}, \tag{2.5}$$

which is an $(n+1) \times (n+1)$ matrix. Additionally we suppose that $\lambda > 0$, $\mathbf{t}$ is an $n \times 1$ column vector with non-negative entries, $\boldsymbol{\alpha}_0$ is an $n \times 1$ column vector with entries that sum up to one and $\mathbf{T}$ is an $n \times n$ matrix with non-negative off-diagonal entries. The column vector $\mathbf{t}$ and the matrix $\mathbf{T}$ are such that each row of $\mathcal{Q}$

sums up to one, alternatively we can write $\mathbf{t} = -\mathbf{T}\mathbf{1}$. On the event $\{\mathcal{J}(\cdot) = 1\}$ the process $X(\cdot)$ decreases linearly with rate $r_1 < 0$ and on the event $\{\mathcal{J}(\cdot) = i\}$, for $i = 2, \ldots, n+1$, $X(\cdot)$ increases linearly with rate $r_i > 0$. Such a MAFP decreases linearly with rate r_1 during OFF-times, which are exponentially distributed with parameter λ, and increases linearly with rates r_i during ON-times, which have a phase-type $(n, \boldsymbol{\alpha}_0, \mathbf{T})$ distribution. This model is motivated by finite capacity systems with an alternating source: during OFF times work is being served with rate r_1 while during ON times work accumulates with rates $r_2, \ldots, r_{n+1}$.

The workload process $\{Q(t)\}_{t\geq 0}$ we are interested in is formally defined as a solution to a two-sided Skorokhod problem, i.e., for a Markov-additive fluid process $\{X(t)\}_{t\geq 0}$, we have

$$Q(t) = Q(0) + X(t) + L(t) - \bar{L}(t). \tag{2.6}$$

In the above expression $\{L(t)\}_{t\geq 0}$ represents the local time at the infimum and $\{\bar{L}(t)\}_{t\geq 0}$ the local time at K. It is known that such a triplet exists and is unique, see [10]. For more details we refer to [8] and references therein. For notational simplicity we assume that $Q(0) = \tau$ and that $\mathcal{J}(0) = 1$, i.e., we start with an OFF time; the cases $\{Q(0) < \tau, \mathcal{J}(0) \neq 1\}$ and $\{Q(0) > \tau, \mathcal{J}(0) \neq 1\}$ can be dealt with analogously at the expense of more complicated expressions. For the MAFP described above the matrix exponent is an $(n+1) \times (n+1)$ matrix. For $q > 0$, denote by $\rho_1(q), \ldots, \rho_{n+1}(q)$ the $n+1$ roots of the equation $\det(\mathcal{Q} - z\boldsymbol{\Delta}_r - q\mathbf{I}) = 0$ and consider, for $k = 1, \ldots, n+1$, the vectors $\mathbf{h}_k(q) = (\mathrm{h}_{k,1}(q), \ldots, \mathrm{h}_{k,n+1}(q))$ defined by

$$\mathrm{h}_{k,1}(q) = 1 \ \ \forall k = 1, \ldots, n+1 \quad \text{and} \quad \mathrm{h}_{k,j}(q) = -\mathbf{e}_{j-1}^{\mathrm{T}}(\mathbf{T} - \rho_k(q)\bar{\boldsymbol{\Delta}}_r - q\mathbf{I})^{-1}\mathbf{t} \ \text{ for } \ j = 2, \ldots, n+1, \tag{2.7}$$

where $\mathbf{e}_j$ is the $n \times 1$ unit column vector with 1 at position j, $\mathbf{T}$ and $\mathbf{t}$ are as in (2.5) and $\bar{\boldsymbol{\Delta}}_r$ is the $n \times n$ diagonal submatrix of $\bar{\boldsymbol{\Delta}}_r$ with r_j at position $(j-1, j-1)$, for $j = 2, \ldots, n+1$. For the vectors defined in (2.7) we have that $(\mathcal{Q} - \rho_k(q)\boldsymbol{\Delta}_r - q\mathbf{I})\mathbf{h}_k(q) = 0$ for all $k = 1, \ldots, n+1$.

3. Result

3.1. The Markov additive fluid process

For the analysis of the occupation time $\alpha(\cdot)$ we observe that the workload process $\{Q(t)\}_{t\geq 0}$ alternates between the two sets $[0, \tau]$ and $(\tau, K]$. Due to the definition of $\{X(t)\}_{t\geq 0}$ both upcrossings and downcrossings of level τ occur with equality. Moreover, we see that an upcrossing of level τ can occur only when the modulating Markov process is in one of the states $2, \ldots, n+1$. Similarly, a downcrossing of level τ can occur only while the modulating Markov process is in state 1. We define the following first passage times, for $\tau \geq 0$,

$$\sigma := \inf_{t>0}\{t : Q(t) = \tau \,|\, Q(0) = \tau, \mathcal{J}(0) = 1\}, \quad T := \inf_{t>0}\{t : Q(t) = \tau \,|\, Q(0) = \tau, \mathcal{J}(0) \neq 1\}.$$

We use the notation $(T_i)_{i\in\mathbb{N}}$ for the sequence of successive downcrossings and $(\sigma_i)_{i\in\mathbb{N}}$ for the sequence of successive upcrossings of level τ. An extension of [7, Thms. 1 and 2] for the case of doubly reflected processes shows that $(T_i)_{i\in\mathbb{N}}$ is a renewal process, and hence the successive sojourn times, $D_1 := \sigma_1$, $D_i := \sigma_i - T_{i-1}$, for $i \geq 2$, and $U_i := T_i - \sigma_i$, for $i \geq 1$, are sequences of well defined random variables. In addition, D_{i+1} is independent of U_i while in general D_i and U_i are dependent. We observe that the random vectors $(D_i, U_i)_{i\in\mathbb{N}}$ are i.i.d. and distributed as a generic random vector (D, U). The double transform of the occupation time $\alpha(\cdot)$ in terms of the joint transform of D and U is given in [16, Theorem 3.1] which we now restate:

Theorem 3.1. *For the transform of the occupation time $\alpha(\cdot)$, and for $q \geq 0, \theta \geq 0$, we have*

$$\int_0^\infty e^{-qt}\,\mathbb{E}\,e^{-\theta\alpha(t)}\mathrm{d}t = \frac{1}{1 - L_{1,2}(q+\theta, q)}\left[\frac{1 - L_1(q+\theta)}{q+\theta} + \frac{L_1(q+\theta) - L_{1,2}(q+\theta, q)}{q}\right], \tag{3.1}$$

where, for $\theta_1, \theta_2 \geq 0$, $L_{1,2}(\theta_1, \theta_2) = \mathbb{E}\, e^{-\theta_1 D - \theta_2 U}$ *and* $L_1(\theta_1) = L_{1,2}(\theta_1, 0) = \mathbb{E}\, e^{-\theta_1 D}$.

Remark 1. *Theorem 3.1 holds when the time instances* $T_m = \sum_{i=1}^{m}(D_i + U_i), m = 1, 2, \ldots$ *are regeneration epochs, i.e. the pairs* $(D_i, U_i)_{i\geq 1}$ *form a sequence of i.i.d. random vectors. This property does not hold in general for an arbitrary MAFP; it does hold for the MAFP presented in Section 2.2 as there is only one state with negative drift.*

To analyze the occupation time it thus suffices to determine the joint transform of the random variables D and U, i.e. $L_{1,2}(\cdot,\cdot)$. We note that the Laplace transforms of the random variables D and U have been derived in [6] for MAFPs. Our proofs can be trivially extended in order to derive the joint transform of the random variables D and U for a general MAFP, i.e. with multiple states with negative drifts. As our interest is in the occupation time $\alpha(\cdot)$, we restrict ourselves to the MAFP considered in Section 2.2 (see Remark 1). Considering the event E_i that an upcrossing of level τ occurs while the modulating process $\mathcal{J}(\cdot)$ is in state i, for $i = 2, \ldots, n+1$, we obtain, for $\theta_1, \theta_2 \geq 0$,

$$\mathbb{E}\left[e^{-\theta_1 D - \theta_2 U}\right] = \mathbb{E}\left[e^{-\theta_1 \sigma - \theta_2 T}\right] = \sum_{i=2}^{n+1} \mathbb{E}\left[e^{-\theta_1 \sigma} 1_{\{E_i\}}\right] \mathbb{E}\left[e^{-\theta_2 T} | E_i\right]. \tag{3.2}$$

In what follows we use, for $\theta_1 \geq 0$ and $i = 2, \ldots, n+1$, the notation

$$z_i(\theta_1) := \mathbb{E}\left[e^{-\theta_1 \sigma} 1_{\{E_i\}}\right] = \mathbb{E}\left[e^{-\theta_1 \sigma} 1_{\{\mathcal{J}(\sigma)=i\}}\right] \quad \text{and} \quad w_i(\theta_2) := \left[e^{-\theta_2 T} | E_i\right] = \mathbb{E}\left[e^{-\theta_2 T} | \, \mathcal{J}(0) = i\right], \tag{3.3}$$

where the last equality follows from the Markov property of the MAFP. It will be shown that these terms can be computed as the solutions of two systems of linear equations. The idea of conditioning on the phase when an *upcrossing* occurs and using the *conditional independence* of the corresponding time epochs has been developed in [3]. The factors appearing in each term in (3.2) can also be determined using the results of [6]. Determining the factors involved in the terms presented above is the main contribution of the analysis that follows. We first present the exact expression for the double transform of the random variables (D, U).

Theorem 3.2. *For* $\theta_1, \theta_2 \geq 0$, *the joint transform of the random variables* D *and* U *is given by*

$$\mathbb{E}\, e^{-\theta_1 D - \theta_2 U} = \frac{1}{C(\theta_2)} \sum_{i=2}^{n+1} z_i(\theta_1) \sum_{k=1}^{n+1} (-1)^{k+1} c_k(\theta_2) \mathrm{h}_{k,i}(\theta_2), \tag{3.4}$$

where the $z_i(\theta_1)$ *for* $i = 2, \ldots, n+1$ *are determined as the solution of a system of linear equations; this system is given below in (3.7); the quantities* $c_k(\theta_2), k = 1, \ldots, n+1$ *and* $C(\theta_2)$ *depend only on* θ_2 *and are defined below in (3.12). The column vectors* $\mathbf{h}_k(\cdot)$ *for* $k = 1, \ldots, n+1$ *are defined in (2.7).*

The outer sum in (3.4) ranging from 2 to $n+1$ represents the conditioning on one of the n phases of the modulating Markov process when an upcrossing occurs, that is the event $\{\mathcal{J}(\sigma) = i\}$, $i = 2, \ldots, n+1$. Observe that an upcrossing of level τ is not possible when $\mathcal{J}(\cdot)$ is in state 1 because then the process $X(\cdot)$ decreases. The inner sum in (3.4) concerns the transforms of T conditional on the event $\{\mathcal{J}(\sigma) = i\}$.

Remark 2. *Observe that, by considering the process* $\{K - Q(t)\}_{t\geq 0}$, *the results can be directly applied to fluid queues with a single state where work accumulates. The same holds for ON-OFF sources with phase-type OFF times and exponential ON times and doubly reflected risk reserve processes with negative phase-type jumps (see also Subsection 3.2).*

Proof of Theorem 3.2. Below we analyze the two expectations in the right-hand side of (3.2) separately.

◦ We determine $z_i(\theta_1)$, for $i = 2, \ldots, n+1$, as the solution of a system of linear equations; this idea was initially developed in [3, Section 5] and essentially relies on the Kella-Whitt martingale for Markov additive processes. The Kella-Whitt martingale for a Markov additive process reflected at the infimum, has, for all $z \geq 0$, $\theta_1 \geq 0$ and for $t \geq 0$, the following form

$$\mathbf{M}(z,t) = \int_0^t e^{-zQ(s)-\theta_1 s}\mathbf{e}_{\mathcal{J}(s)}\mathrm{d}s(\mathcal{Q} - z\mathbf{\Delta}_r - \theta_1\mathbf{I}) + e^{-z\tau}\mathbf{e}_{\mathcal{J}(0)} - e^{-zQ(t)-\theta_1 t}\mathbf{e}_{\mathcal{J}(t)} - z\int_0^t e^{-\theta_1 s}\mathbf{e}_{\mathcal{J}(s)}\mathrm{d}L(s). \tag{3.5}$$

The expression above follows from the general form of the Kella-Whitt martingale given in (2.4) by considering the process $Y(\cdot)$ defined by $Y(t) := \tau + L(t) + \theta_1 t/z$, for $t \geq 0$. Furthermore, due to the construction of the model we have that $\mathcal{J}(0) = 1$. Applying the optional sampling theorem for the stopping time σ, we obtain, for all $z \geq 0$,

$$\mathbb{E}\left[\int_0^\sigma e^{-zQ(s)-\theta_1 s}\mathbf{e}_{\mathcal{J}(s)}\mathrm{d}s\right](\mathcal{Q} - z\mathbf{\Delta}_r - \theta_1\mathbf{I}) = e^{-z\tau}\mathbf{z}(\theta_1) - e^{-z\tau}\mathbf{e}_1 + z\boldsymbol{\ell}(\theta_1), \tag{3.6}$$

where

$$\mathbf{z}(\theta_1) = \mathbb{E}\left[e^{-\theta_1\sigma}\mathbf{e}_{\mathcal{J}(\sigma)}\right] = \Big(0, \mathbb{E}\left[e^{-\theta_1\sigma}1_{\{\mathcal{J}(\sigma)=2\}}\right], \ldots, \mathbb{E}\left[e^{-\theta_1\sigma}1_{\{\mathcal{J}(\sigma)=n+1\}}\right]\Big) = \Big(0, z_2(\theta_1), \ldots, z_{n+1}(\theta_2)\Big)$$

and

$$\boldsymbol{\ell}(\theta_1) = \mathbb{E}\left[\int_0^\sigma e^{-\theta_1 s}\mathbf{e}_{\mathcal{J}(s)}\mathrm{d}L(s)\right] = \Big(\mathbb{E}\left[\int_0^\sigma e^{-\theta_1 s}1_{\{\mathcal{J}(s)=1\}}\mathrm{d}L(s)\right], 0, \ldots, 0\Big) = \Big(\ell(\theta_1), 0, \ldots, 0\Big).$$

The row vector $\boldsymbol{\ell}(\theta_1)$ represents the local time at the infimum up to the stopping time σ; the process $Q(\cdot)$ can hit level 0 only on the event $\{\mathcal{J}(s) = 1\}$. Consider the $n+1$ roots $\rho_1(\theta_1), \ldots, \rho_{n+1}(\theta_1)$ of the equation $\det(\mathcal{Q} - z\mathbf{\Delta}_r - q\mathbf{I}) = 0$, and the corresponding column vectors $\mathbf{h}_k(\theta_1)$, for $k = 1, \ldots, n+1$ as defined in (2.7). Substituting $z = \rho_k(\theta_1)$ in (3.6) and taking the inner products with the column vectors $\mathbf{h}_k(\theta)$ we obtain, for $k = 1, \ldots, n+1$, the system of equations

$$\mathbf{z}(\theta_1) \cdot \mathbf{h}_k(\theta_1) + e^{\rho_k(\theta_1)\tau}\rho_k(\theta_1)\ell(\theta_1) = 1. \tag{3.7}$$

Solving this system of equations we obtain the $z_i(\theta_1)$, for $i = 2, \ldots, n+1$, and $\ell(\theta_1)$ as a by-product.

◦ Next, consider the second expectation in each of the summands in the RHS of (3.2), i.e., the term $w_i(\theta_2) = \mathbb{E}\left[e^{-\theta_2 T} \middle| \mathcal{J}(0) = i\right]$. This expectation represents the transform of the first time the process $X(\cdot)$ hits level τ given that $\mathcal{J}(0) = i$, for $i = 2, \ldots, n+1$. The Kella-Whitt martingale for a MAFP reflected at K, has, for all $z \geq 0$, $\theta_2 \geq 0$ and for $t \geq 0$, the following form:

$$\begin{aligned}\mathbf{M}_K(z,t) =& \int_0^t e^{-zQ(s)-\theta_2 s}\mathbf{e}_{\mathcal{J}(s)}\mathrm{d}s(\mathcal{Q} - z\mathbf{\Delta}_r - \theta_2\mathbf{I}) + e^{-z\tau}\mathbf{e}_{\mathcal{J}(0)} - e^{-zQ(t)-\theta_2 t}\mathbf{e}_{\mathcal{J}(t)} \\ &+ ze^{-zK}\int_0^t e^{-\theta_2 s}\mathbf{e}_{\mathcal{J}(s)}\mathrm{d}\bar{L}(s).\end{aligned} \tag{3.8}$$

The expression above follows from the general form of the Kella-Whitt martingale given in (2.4) by considering the process $Y(\cdot)$ defined by $Y(t) := \tau - \bar{L}(t) + \theta_2 t/z$, for $t \geq 0$. A similar argument as for the stopping time σ and (3.3) yields n systems of linear equations; for each $i = 2, \ldots, n+1$, we solve for the unknowns $w_i(\cdot)$ and $\bar{\ell}_j(\cdot)$, $j = 2, \ldots, n+1$, using the following system:

$$w_i(\theta_2) + \sum_{j=2}^{n+1}\bar{\ell}_j(\theta_2)\Big(e^{-\rho_k(\theta_2)(K-\tau)}\rho_k(\theta_2)\mathrm{h}_{k,j}(\theta_2)\Big) = \mathrm{h}_{k,i}(\theta_2) \quad \text{for } k = 1, \ldots, n+1, \tag{3.9}$$

where $\bar{\ell}_j(\theta_2) = \mathbb{E}\left[\int_0^T e^{-\theta_2 s} 1_{\{\mathcal{J}(s)=j\}} \mathrm{d}\bar{L}(s)\right], j = 2, \ldots, n+1$. We define $\mathbf{D}$ as the $(n+1) \times n$ dimensional matrix with elements, for $i = 1, \ldots, n+1$, $j = 1, \ldots, n$,

$$\mathbf{D}(i,j) = \rho_i(\theta_2) \mathrm{h}_{i,j+1}(\theta_2) e^{-\rho_i(\theta_2)(K-\tau)}. \tag{3.10}$$

We denote by $\mathbf{D}_k$ the matrix $\mathbf{D}$ without its k-th row. Using the method of determinants we can write $w_i(\theta_2)$ in the following form:

$$w_i(\theta_2) = \mathbb{E}\left[e^{-\theta_2 T} | \, \mathcal{J}(0) = i\right] = \frac{\sum_{k=1}^{n+1} (-1)^{1+k} c_k(\theta_2) \mathrm{h}_{k,i}(\theta_2)}{C(\theta_2)}, \tag{3.11}$$

where, for $k = 1, \ldots, n+1$,

$$c_k(\theta_2) = \det(\mathbf{D}_k) \quad \text{and} \quad C(\theta_2) = \sum_{k=1}^{n+1} (-1)^{1+k} c_k(\theta_2). \tag{3.12}$$

Substituting the expression found for $w_i(\theta_2)$ in (3.11) into (3.2) yields the result of Theorem 3.2 with the terms $z_i(\theta_2)$, $i = 2, \ldots, n+1$, given by the system of equations in (3.7). □

3.2. The finite buffer queue

Using the result of Theorem 3.2 we can also study the occupation time of the workload process in a ***finite-buffer queue*** with phase-type service time distribution. Consider a queue where customers arrive according to a Poisson process with rate λ and have a phase-type service time distribution with representation $(n, \boldsymbol{\alpha}_0, \mathbf{T})$. Moreover, the queue has finite capacity K and work is served with rate r_1. The workload process $\{Q(t)\}_{t \geq 0}$ is modeled using a reflected compound Poisson process with negative drift $r_1 < 0$ and upward jumps with a phase-type $(n, \boldsymbol{\alpha}_0, \mathbf{T})$ distribution. Such a process has Laplace exponent equal to

$$\phi(s) = -sr_1 - \lambda + \lambda \hat{B}(s) = -sr_1 - \lambda + \lambda \boldsymbol{\alpha}_0^{\mathrm{T}} (s\mathbf{I} - \mathbf{T})^{-1} \mathbf{t}, \quad s \geq 0, \tag{3.13}$$

where $\mathbf{t} = -\mathbf{T}\mathbf{1}$. As for the MAFP in Section 3.1 we determine the joint transform of the random variables D and U. First we introduce some notation. Define, for $k = 1, \ldots, n+1$, the vectors $\mathbf{h}_k(\cdot)$ as:

$$\mathrm{h}_{k,1}(\cdot) = 1, \;\; \forall k = 1, \ldots, n+1 \;\text{ and }\; \mathrm{h}_{k,j}(\cdot) = \hat{B}_j(p_k(\cdot)), \;\; j = 2, \ldots, n+1, \tag{3.14}$$

where $p_k(q), k = 1, \ldots, n+1$, are the $n+1$ roots of the equation $\phi(s) = q$. Consider the following system of linear equations

$$\sum_{j=2}^{n+1} z_j(\theta_1) \mathrm{h}_{k,j}(\theta_1) + p_k(\theta_1) e^{p_k(\theta_1)\tau} \ell(\theta_1) = 1, \;\; k = 1, \ldots, n+1, \tag{3.15}$$

and define $c_k(\cdot), k = 1, \ldots, n+1$ and $C(\cdot)$ as in (3.12) above with the defference that $\rho_k(\cdot)$ is replaced by $p_k(\cdot)$.

Corollary 3.1. *Consider a compound Poisson process with negative drift $r_1 < 0$ and upward jumps with a phase-type $(n, \boldsymbol{\alpha}_0, \mathbf{T})$ distribution. Consider the process reflected at the infimum and at level $K > 0$. For $\theta_1, \theta_2 \geq 0$, the joint transform of the random variables D and U is given by*

$$\mathbb{E}\, e^{-\theta_1 D - \theta_2 U} = \frac{1}{C(\theta_2)} \sum_{i=2}^{n+1} z_i(\theta_1) \sum_{k=1}^{n+1} (-1)^{k+1} c_k(\theta_2) \mathrm{h}_{k,i}(\theta_2), \tag{3.16}$$

where $c_k(\theta_2), k = 1, \ldots, n+1$ and $C(\theta_2)$ are as above; $z_i(\theta_1)$ for $i = 2, \ldots, n+1$ are determined as the solution of (3.15) and $\mathbf{h}_k(\cdot), k = 1, \ldots, n+1$ is as in (3.14).

The workload process $\{Q(t)\}_{t\geq 0}$ in the finite-buffer queue can be studied as the limit of a MAFP in the following sense, see also [4, Section 7]. Following the construction presented in Section 2.2 we define, for $r > 0$, the MAFP $\{X^r(t), \mathcal{J}^r(t)\}_{t\geq 0}$ where the Markov process has state space $E = \{1, \ldots, n+1\}$ and generator $\mathcal{Q}^r$ given by

$$\mathcal{Q}^r = \begin{bmatrix} -\lambda & \lambda\alpha_{\mathbf{0}}^{\mathrm{T}} \\ r\mathbf{t} & r\mathbf{T} \end{bmatrix},$$

which is a $(n+1)\times(n+1)$ matrix. We also let the positive rates be equal, i.e, $r_2 = \ldots = r_{n+1} = r$ and we send $r \to \infty$ later on. The assumptions on $\lambda, \mathbf{t}, \boldsymbol{\alpha}_0$ and $\mathbf{T}$ are the same as in Section 2.2. On the event $\{\mathcal{J}^r(\cdot) = 1\}$ the process $X^r(\cdot)$ decreases with rate $r_1 < 0$ and on the event $\{\mathcal{J}^r(\cdot) = i\}$, for $i = 2, \ldots, n+1$, the process $X^r(\cdot)$ increases with rate $r > 0$. Letting $r \to \infty$ the process $(X^r(t), \mathcal{J}^r(t))_{t\geq 0}$ converges path-wise to a compound Poisson process with linear rate $r_1 < 0$ and jumps in the upward direction with phase-type $(n, \boldsymbol{\alpha}_0, \mathbf{T})$ distribution. The workload process $\{Q^r(t)\}_{t\geq 0}$ converges to $\{Q(t)\}_{t\geq 0}$, i.e. a reflected compound Poisson process, which follows by the continuity of the reflection operators with respect to the D_1 topology. Hence the joint transform of D and U is computed by using the result established in Theorem 3.2 and letting $r \to \infty$.

4. Numerical Computation

In this section we describe how to numerically evaluate the distribution function (or the density) of the occupation time. Essential is the joint transform of the consecutive periods below and above τ, i.e., $\mathbb{E}\, e^{-\theta_1 D - \theta_2 U}$ for $\theta_1 \geq 0, \theta_2 \geq 0$. It enables the computation of the double transform of the occupation time, which in turn uniquely characterizes the distribution of this occupation time. The idea is to evaluate $\mathbb{P}(\alpha(t) \leq s)$ by numerically inverting the double transform using the methodology of [1]. The methodology we present is rather straightforward to implement and yields a highly accurate numerical approximation for the distribution function of the occupation time. We first present an algorithm which uses Theorem 3.4 in order to evaluate the distribution function (or the density function) of the occupation time $\alpha(\cdot)$. Then, using the technique presented above, we can let $r \to \infty$ to obtain the distribution function of the occupation time $\alpha(\cdot)$ in the M/G/1 queue with a phase-type service distribution.

Algorithm:. **Input:** $t \geq 0, s \in [0, \infty)$ **Output:** The distribution function $\mathbb{P}(\alpha(t) \leq s)$ (or the density).

(1) Compute $L_{1,2}(\theta_1, \theta_2) = \mathbb{E}\, e^{-\theta_1 D - \theta_2 U}$ and $L_1(\theta) = L_{1,2}(\theta, 0)$ using Theorem 3.2 and by solving the systems of linear equations given in (3.7) and (3.9).

(2) Compute the double transform of the occupation time $\alpha(\cdot)$ using Theorem 3.1.

(3) Use Laplace inversion in order to compute $\mathbb{P}(\alpha(t) \leq s)$ (or the density).

We first numerically compute the right hand sides of (3.4) and we obtain a highly accurate numerical approximation of the joint transform $\mathbb{E}\, e^{-\theta_1 D - \theta_2 U}, \theta_1 \geq 0, \theta_2 \geq 0$. We then use (3.1) and the Laplace inversion techniques presented in [1] to compute the density function of the occupation time. We used the Euler-Euler algorithm with $M = 10$. In Figure 1 below we show the density function of the occupation time $\alpha(\cdot)$ in a MAFP-driven queue (left panel) and a M/G/1 queue (right panel). For both cases we consider a time horizon of $t = 100$ time units. In the M/G/1 the arrival rate is equal to $\lambda = 1.05$, whereas in the MAFP the OFF-time is exponential with parameter $\lambda = 1.05$. The depletion rate has been normalized to 1 in the MAFP model (i.e., $r_1 = -1$). The parameters of the two models are chosen in such a way so that the *load* in the system is the same for all cases and equal to $\rho = 0.945$; we define this load in the M/G/1 model in the natural way (i.e., as

the product of the arrival rate and the mean jump size), and in the MAFP model as the product of λ and the mean increase of the fluid level during the ON-time (for the case that there would not have been a finite buffer capacity). We have chosen $\tau = 0.8$ and $K = 2$.

We consider MAFPs with ON-times having an Erlang, exponential, and Coxian distribution, respectively (where we recall that these have a coefficient of variation less than one, equal to one, and greater than one, respectively). For the exponential distribution we choose $\mu = 2$ and $r_2 = 1.8$, such that $\rho = 1.05 \cdot 1.8/2 = 0.945$, as desired. For the Erlang case we take two phases, corresponding with states 2 and 3 of the Markov process $\mathcal{J}(\cdot)$, both having an exponentially distributed duration with parameter $\mu_2 = \mu_3 = 6$, where the buffer content increases at rates $r_2 = 1.8$ and $r_3 = 3.6$; it is easily checked that the load is $\rho = 0.945$. For the Coxian case, with two phases (again corresponding with states 2 and 3), we have with probability $p = 0.5$ an ON-time which is distributed as the sum of two exponential distributions with parameters $\mu_2 = 18, \mu_3 = 2.25$ and otherwise an exponentially distributed ON-time with parameter $\mu_3 = 2.25$; with again $r_2 = 1.8$ and $r_3 = 3.6$, it is easily checked that the load is $\rho = 0.945$.

For the M/G/1 queue we consider the cases the jump size has an Erlang, exponential, and Coxian distribution. (A) The exponential distribution has parameter $\frac{10}{9} = 1.111$. (B) For the Erlang distribution we consider $m = 2$ (where each phase is exponential with parameter 2.222) as well as $m = 4$ (where each phase is exponential with parameter 4.444). (C) For the Coxian distribution we take $m = 2$ (i.e., 2 phases), $p = 0.5$ and the exponential phases have parameters 5.555 and 0.694, respectively. It is easily verified that all these settings lead to $\rho = 0.945$. We have also included an Erlang(4) jump distribution, again chosen such that $\rho = 0.945$.

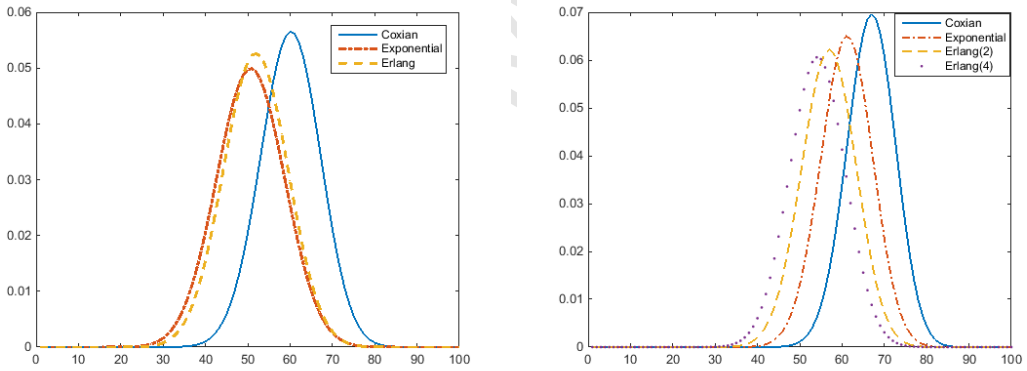

Figure 1: Distribution function and density of the occupation time

In the left panel of Figure 2 we consider the M/G/1 queue in which we vary the arrival rate, so as to study the occupation time for different values of ρ. The service time has an Erlang distribution with two phases (each exponentially distributed with parameter 2.222). In the right panel we vary the buffer capacity K; the process we considered was the MAFP with OFF-times having a Coxian distribution. The parameters of the distribution are the same as those that we used above.

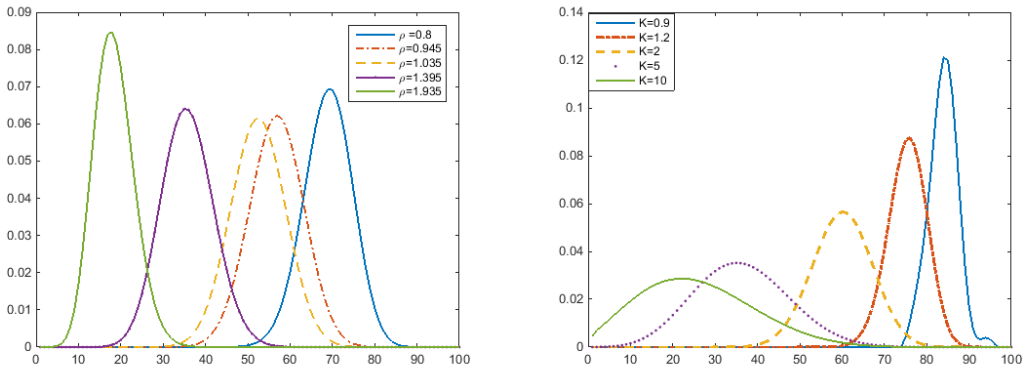

Figure 2: Density function of the occupation time as ρ and K vary

Acknowledgements

We would like to thank the associate editor for his inspiring comments and the anonymous referee who provided us with thoughtful comments. Their suggestions helped us improve the paper significantly. The research of N. Starreveld and M. Mandjes is partly funded by the NWO Gravitation project NETWORKS, grant number 024.002.003.

[1] J. ABATE AND W. WHITT (2006). *A unified framework for numerically inverting Laplace transforms.* INFORMS Journal on Computing, Vol. 18, pp. 408-421.

[2] S. ASMUSSEN (2003). *Applied Probability and Queues,* 2nd edition. Springer, New York.

[3] S. ASMUSSEN (2014). *Lévy processes, phase-type distributions and martingales.* Stochastic Models, Vol. 30, pp. 443-468.

[4] S. ASMUSSEN AND O. KELLA (2000). *A Multi-dimensional Martingale for Markov Additive Processes and its Applications.* Advances in Applied Probability, Vol. 32, No. 2, pp. 376-393.

[5] O. BARON AND J. MILNER (2009). *Staffing to maximize profit for call centers with alternate service-level agreements.* Operations Research, Vol. 57, pp. 685-700.

[6] N. BEAN, M. O'REILLY AND P. TAYLOR (2009). *Hitting probabilities and hitting times for stochastic fluid flows: The bounded model.* Probability in the Engineering and Informational Sciences, 23(1), 121-147.

[7] J. COHEN AND M. RUBINOVITCH (1977). *On level crossings and cycles in dam processes.* Mathematics of Operations Research, Vol. 2, pp. 297-310.

[8] K. DEBICKI AND M. MANDJES (2015). *Queues and Lévy Fluctuation Theory.* Springer, New York.

[9] J. IVANOVS, O. BOXMA AND M. MANDJES (2010). *Singularities of the matrix exponent of a Markov additive process with one-sided jumps.* Stochastic Processes and their Applications, Vol. 120, Issue 9, pp. 1776–1794.

[10] L. KRUK, J. LEHOCZKY, K. RAMANAN AND S. SHREVE (2007). *An explicit formula for the Skorokhod map on* $[0, \alpha]$. The Annals of Probability, Vol. 35, No. 5, 1740-1768.

[11] A. Kyprianou, J. Pardo and J. Pérez (2014). *Occupation times of refracted Lévy processes*. Journal of Theoretical Probability, Vol. 27, pp. 1292-1315.

[12] D. Landriault, J. Renaud and X. Zhou (2011). *Occupation times of spectrally negative Lévy processes with applications*. Stochastic Processes and their Applications, Vol. 121, pp. 2629-2641.

[13] R. Loeffen, J. Renaud and X. Zhou (2014). *Occupation times of intervals until passage times for spectrally negative Lévy processes*. Stochastic Processes and their Applications, Vol. 124, pp. 1408-1435.

[14] A. Roubos, R. Bekker and S. Bhulai (2015). *Occupation times for multi-server queues*. Submitted.

[15] A. Roubos, G.M. Koole and R. Stolletz (2012). *Service-level variability of inbound call centers*. Manufacturing & Service Operations Management, Vol. 14, pp. 402-413.

[16] N. J. Starreveld, R. Bekker and M. Mandjes (2017). *Occupation times for regenerative processes with Lévy applications*. Submitted. arXiv:1602.05131.

[17] N. J. Starreveld, R. Bekker and M. Mandjes (2017). *Occupation times for the finite buffer fluid queue with phase-type ON-times*. arXiv:1703.05500

[18] L. Takács (1957). *On certain sojourn time problems in the theory of stochastic processes*. Acta Mathematica Academiae Scientiarum Hungarica, Vol. 8, pp. 169-191.

[19] S. Zacks (2012). *Distribution of the total time in a mode of an alternating renewal process with applications*. Sequential Analysis, Vol. 31, pp. 397-408.