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

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

March 15, 2017

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. We construct an alternating sequence of sojourn times $D_1, U_1, \ldots$ where the pairs $(D_i, U_i)_{i\in\mathbb{N}}$ are i.i.d. random vectors. We use this sequence to determine the distribution function of the occupation time in terms of its double transform.

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 University 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.

Mathematics Subject Classification: 60J55; 60J75; 60K25.

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 specific cases explicit analysis of the finite-buffer queue *is* possible. 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

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

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 [?, ?] show that there is severe fluctuation in the service level, even when measured over periods of several hours up to a day. 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). The fact that the service level fluctuates on such rather long time scales has been observed in the queueing community only relatively recently (see [?, ?, ?] for some call center and queueing applications). 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 [?, ?, ?, ?]. Another branch is concerned with occupation times of spectrally negative Lévy processes, see e.g. [?, ?, ?]. The results established typically concern occupation times until a first passage time, whereas [?] focuses on refracted Lévy processes. In [?] spectrally positive Lévy processes with reflection at the infimum were studied as a special case; we also refer to [?] and references therein for additional literature. A natural extension of Lévy processes are Markov-modulated Lévy processes; for the case of a Markov-modulated Brownian motion the occupation time has been analyzed in [?]. 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 [?]. 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 [?], 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. Some discussion related to computability can be found in Section 4.

2 Model description and preliminaries

We consider the finite capacity fluid queue with subsequent ON and OFF times. In the fluid queue work arrives from a source with a linear rate and the source switches between two modes, ON and OFF. The ON times correspond to time periods in which the source is active, whereas the OFF times correspond to time periods in which the source is inactive. During the ON times work accumulates at a linear rate depending on the state of an underlying Markov chain while during OFF times work decreases again with a linear rate. From the above it follows that OFF times have an exponential distribution and the ON times have a phase-type

distribution. The workload capacity is K and work that does not fit is rejected; see Subsection **??** 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 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 $\vec{\alpha}_0$ the initial probability distribution of the Markov process, by $\mathbf{T}$ the *phase generator*, i.e., the $|E| \times |E|$ rate matrix between the transient states and by $\vec{t}$ the *exit vector*, i.e., the $|E|$- dimensional rate vector between the transient states and the absorbing state ∂. The vector $\vec{t}$ 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 $(n, \vec{\alpha}_0, \mathbf{T})$ where $|E| = n$. The cardinality of the state space E, i.e., n, represents the *number of phases* of the phase-type distribution B; for simplicity we assume that $E = \{1, \ldots, n\}$. In what follows we denote by B a phase-type distribution with representation $(n, \vec{\alpha}_0, \mathbf{T})$; for a phase-type distribution with representation $(n, \vec{e}_i, \mathbf{T})$ 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 [**?**, Thm. 4.2]. For a phase-type distribution with representation $(n, \vec{\alpha}_0, \mathbf{T})$, the cumulative distribution function $B(\cdot)$, the density $b(\cdot)$ and the Laplace transform $\hat{B}[\cdot]$ are given in [**?**, Prop. 4.1]. In particular, for $x \geq 0$ and $s \geq 0$, we have

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

When the phase-type distribution has representation $(n, \vec{e}_i, \mathbf{T})$ 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 [**?**, **?**] 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 [**?**, 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 $\mathcal{J}(\cdot)$. Letting $\{T_i, i \geq 0\}$ be the jump epochs of the Markov process $\mathcal{J}(\cdot)$ (with $T_0 = 0$) we obtain the following expression for the process $X(\cdot)$,

$$\begin{aligned} X(t) = X_0 &+ \sum_{m\geq 1}\sum_{1\leq i\leq n} r_i\,(T_m - T_{m-1})\,1_{\{\mathcal{J}(T_{m-1})=i, T_m \leq t\}} \\ &+ \sum_{m\geq 1}\sum_{1\leq i\leq n} r_i(t - T_{m-1})1_{\{\mathcal{J}(T_{m-1})=i, T_{m-1}\leq t < T_m\}}, \quad t \geq 0 \end{aligned} \tag{2.2}$$

where $X_0 \in \mathcal{F}_0$ and is independent of the Markov process $\mathcal{J}(\cdot)$ and the rates $r_1, \ldots, r_n$. The process $X(\cdot)$ defined in (**??**) will be referred to as a *Markov-additive fluid process* and abbreviated as MAFP. For $\alpha \in \mathbb{C}^{\mathrm{Re}\geq 0}$, the *matrix exponent* of the MAFP is defined as

$$F(\alpha) = \mathcal{Q} + \alpha\ \mathrm{diag}(r_1, \ldots, r_n) = \mathcal{Q} + \alpha\mathbf{\Delta}_r. \tag{2.3}$$

In what follows we shall need information concerning the roots of the equation $\det(F(\alpha) - q\mathbf{I}) = \det(\mathcal{Q} + \alpha\mathbf{\Delta}_r - q\mathbf{I}) = 0$, where $\mathbf{\Delta}_r = \mathrm{diag}(r_1, \ldots, r_n)$ and $q \geq 0$. From [**?**] we have that there exist n values $\rho_1(q), \ldots, \rho_n(q)$

and corresponding vectors $\vec{h}_1(q), \ldots, \vec{h}_n(q)$ such that, for each $k = 1, \ldots, n$, $\det(\mathcal{Q} + \rho_k(q)\mathbf{\Delta}_r - q\mathbf{I}) = 0$ and $(\mathcal{Q} + \rho_k(q)\mathbf{\Delta}_r - q\mathbf{I})\vec{h}_k(q) = 0$.

The Kella-Whitt martingale The counterpart of the Kella-Whitt martingale for *Markov-additive processes* was established in [**?**]; 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 $\alpha \in \mathbb{C}^{\mathrm{Re}\geq 0}$. Then, for every initial distribution $(X(0), \mathcal{J}(0))$,

$$M(\alpha, t) := \int_0^t e^{\alpha Z(s)} \vec{e}_{\mathcal{J}(s)} \mathrm{d}s F(\alpha) + e^{\alpha Z(0)} \vec{e}_{\mathcal{J}(0)} - e^{\alpha Z(t)} \vec{e}_{\mathcal{J}(t)} + \alpha \int_0^t e^{\alpha Z(s)} \vec{e}_{\mathcal{J}(s)} \mathrm{d}Y(s) \tag{2.4}$$

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

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

which is a $(n+1) \times (n+1)$ matrix. Additionally we suppose that $\lambda > 0$, $\vec{t}$ is a $n \times 1$ column vector with negative entries, $\vec{\alpha}_0$ is a $n \times 1$ column vector with entries that sum up to one and $\mathbf{T}$ is a $n \times n$ matrix with non-negative entries. The vector $\vec{t}$ and the matrix $\mathbf{T}$ are such that each row of $\mathcal{Q}$ sums up to one, alternatively we can write $\vec{t} = -\mathbf{T}\mathbf{1}$. 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 $(n, \vec{\alpha}_0, \mathbf{T})$ distribution. Depending on the state of the modulating process we have a different rate. 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 $\{X(t)\}_{t\geq 0}$ as defined in (**??**), 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. Informally, for $t > 0$, $L(t)$ is the amount that has to be added to $X(t)$ so that it stays non-negative while $\bar{L}(t)$ is the amount that has to be subtracted from $X(t) + L(t)$ so that it stays below level K. It is known that such a triplet exists and is unique, see [**?**, **?**]. For more details we refer to [**?**] 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 a $(n+1) \times (n+1)$ matrix. For $q > 0$, denote by $\rho_1(q), \ldots, \rho_{n+1}(q)$ the $n+1$ roots of the equation $\det(\mathcal{Q} + \alpha\mathbf{\Delta}_r - q\mathbf{I}) = 0$ and consider, for $k = 1, \ldots, n+1$, the vectors $\vec{h}_k(q) = (h_{k,1}(q), \ldots, h_{k,n+1}(q))$ defined by

$$h_{k,1}(q) = 1 \ \ \forall k = 1, \ldots, n+1 \quad \text{and} \quad h_{k,j}(q) = -\vec{e}_j^{\intercal}(\mathbf{T} + \rho_k(q)\mathbf{\Delta}_r - q\mathbf{I})^{-1}\vec{t} \ \text{ for } \ j = 2, \ldots, n+1, \tag{2.7}$$

where $\vec{e}_j$ is the unit column vector with 1 at position j, $\mathbf{T}$ and $\vec{t}$ are as in (??). For the vectors defined in (??) we have that $(\mathcal{Q}+\rho_k(q)\mathbf{\Delta}_r - q\mathbf{I})\vec{h}_k(q) = 0$ 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 [?, 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). In Figure ?? a realization of $Q(\cdot)$ is depicted.

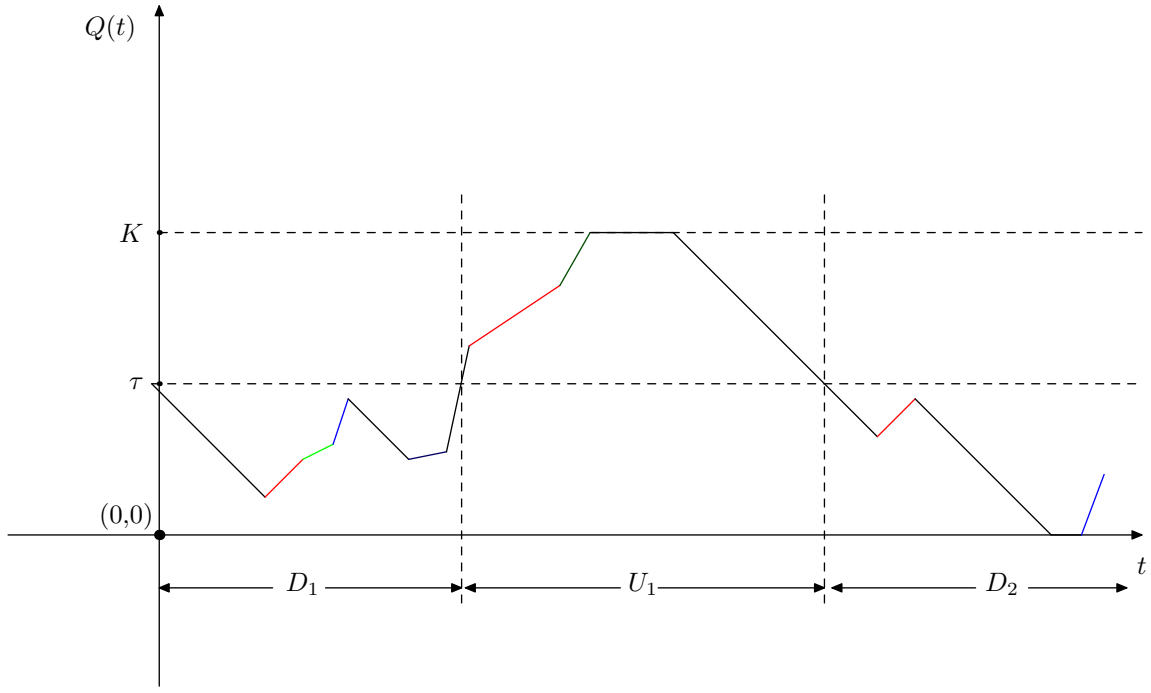

Figure 1: The workload process in a finite capacity fluid queue

The double transform of the occupation time $\alpha(\cdot)$ in terms of the joint distribution of U and D is given in [?, 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*

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

where, for $\theta_1, \theta_2 \geq 0$,

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

To analyze the occupation time it thus suffices to determine the joint transform of the random variables D and U, i.e., $L_{1,2}(\cdot,\cdot)$. For $i = 2,\ldots,n+1$ we define the first hitting time of level τ with initial condition $(X(0),\mathcal{J}(0)) = (\tau,i)$ as follows:

$$T_i := \inf_{t\geq 0}\{t : Q(t) = \tau | Q(0) = \tau, \mathcal{J}(0) = i\} \quad \text{and} \quad w_i(\theta_2) = \mathbb{E}\left[e^{-\theta_2 T_i}\right] = \mathbb{E}\left[e^{-\theta_2 T}|\,\mathcal{J}(0) = i\right].$$

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 $\theta_1,\theta_2 \geq 0$,

$$\mathbb{E}\left[e^{-\theta_1 D-\theta_2 U}\right] = \mathbb{E}\left[e^{-\theta_1\sigma-\theta_2 T}\right] = \sum_{i=2}^{n+1}\mathbb{E}\left[e^{-\theta_1\sigma}1_{\{E_i\}}\right]\mathbb{E}\left[e^{-\theta_2 T}|E_i\right] = \sum_{i=2}^{n+1}\mathbb{E}\left[e^{-\theta_1\sigma}1_{\{E_i\}}\right]\mathbb{E}\left[e^{-\theta_2 T_i}\right]. \tag{3.1}$$

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

$$z_i(\theta_1) := \mathbb{E}\left[e^{-\theta_1\sigma}1_{\{E_i\}}\right] = \mathbb{E}\left[e^{-\theta_1\sigma}1_{\{\mathcal{J}(\sigma)=i\}}\right]. \tag{3.2}$$

It will be shown that these terms can be computed as the solution of a system 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 [**?**]. Determining the factors involved in the terms presented above is the main contribution of the analysis that follows. We now 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*

$$\mathbb{E}\,e^{-\theta_1 D-\theta_2 U} = \frac{1}{C(\theta_2)}\sum_{i=2}^{n+1} z_i(\theta_1)\sum_{j=1}^{n+1}(-1)^{j+1}c_j(\theta_2)e^{-\rho_j(\theta_2)(K-\tau)}h_{j,i}(\theta_2), \tag{3.3}$$

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

The outer sum in (**??**) ranging from 2 to $n+1$ represents the conditioning on one of the $n+1$ 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 terms $z_i(\theta_1)$, as defined in (**??**), denote the transforms of σ on the event the upcrossing of level τ occurs while the modulating Markov process is in state i, and the inner sum in (**??**) concerns the transform of T conditional on the event $\{\mathcal{J}(\sigma) = i\}$. The Markov property of the workload process yields that

$$\mathbb{E}\left[e^{-\theta_2 T}|E_i\right] = \mathbb{E}\left[e^{-\theta_2 T}|\,\mathcal{J}(\sigma) = i\right] = \mathbb{E}\left[e^{-\theta_2 T}|\,\mathcal{J}(0) = i\right] = w_i(\theta_2). \tag{3.4}$$

Proof of Theorem **??**. The proof relies on the decomposition given in (**??**). Below we analyze the two expectations at the RHS of (**??**) 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 [**?**, 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 $\alpha \geq 0$, $\theta_1 \geq 0$ and for $t \geq 0$, the following form

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

The expression above follows from the general form of the Kella-Whitt martingale given in (**??**) by considering the process $Y(\cdot)$ defined by $Y(t) := \tau + L(t) - \theta_1 t/\alpha$, for $t \geq 0$. This gives $Z(t) = \tau + X(t) + L(t) - \theta_1 t/\alpha = Q(t) - \theta_1 t/\alpha$. The process $Y(\cdot)$ has paths of bounded variation and is also continuous since the local time at the infimum $L(\cdot)$ is a continuous process. Hence, $M(\alpha, \cdot)$ is a zero-mean martingale. 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 $\alpha \geq 0$,

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

where

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

and

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

The vector $\vec{\ell}(\theta_1)$ 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 of the equation $\det(\mathcal{Q} + \alpha \mathbf{\Delta}_r - \theta_1 \mathbf{I}) = 0$, denoted by $\rho_1(\theta_1), \ldots, \rho_{n+1}(\theta_1)$, and the corresponding vectors $\vec{h}_k(\theta_1)$, for $k = 1, \ldots, n+1$ as defined in (**??**). Substituting $\alpha = \rho_k(\theta_1)$ in (**??**) and taking the inner products with the vectors $\vec{h}_k(\theta)$ we obtain, for $k = 1, \ldots, n+1$, the system of equations

$$e^{\rho_k(\theta_1)\tau} \vec{z}(\theta_1)^\intercal \cdot \vec{h}_k(\theta_1) - e^{\rho_k(\theta_1)\tau} \vec{e}_1^\intercal \cdot \vec{h}_k(\theta_1) - \rho_k(\theta_1) \vec{\ell}(\theta_1)^\intercal \cdot \vec{h}_k(\theta_1) = 0.$$

Hence we obtain the following system of $n+1$ linear equations and $n+1$ unknowns, i.e., $z_i(\theta_1)$ for $i = 2, \ldots, n+1$ and $\ell(\theta_1)$,

$$e^{\rho_k(\theta_1)\tau} \sum_{j=2}^{n+1} z_j(\theta_1) h_{k,j}(\theta_1) - e^{\rho_k(\theta_1)\tau} - \rho_k(\theta_1) \ell(\theta_1) = 0, \quad k = 1, \ldots, n+1. \tag{3.7}$$

Solving this system of equations we obtain the $z_i(\theta_1)$, for $i = 2, \ldots, n+1$.

◦ Next, consider the second expectation in each of the summands at the RHS of (**??**), i.e., the term $w_i(\theta_2) = \mathbb{E}\left[e^{-\theta_2 T} | \, \mathcal{J}(0) = i\right]$. 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 $\alpha \geq 0, \theta_2 \geq 0$ and for $t \geq 0$, the following form:

$$M_K(\alpha, t) = \int_0^t e^{\alpha Q(s) - \theta_2 s} \vec{e}_{\mathcal{J}(s)} \mathrm{d}s (\mathcal{Q} + \alpha \mathbf{\Delta}_r - \theta_2 \mathbf{I}) + e^{\alpha \tau} \vec{e}_{\mathcal{J}(0)} - e^{\alpha Q(t) - \theta_2 t} \vec{e}_{\mathcal{J}(t)} - \alpha e^{\alpha K} \int_0^t e^{-\theta_2 s} \vec{e}_{\mathcal{J}(s)} \mathrm{d}\bar{L}(s).$$

The expression above follows from the general form of the Kella-Whitt martingale given in (**??**) by considering the process $Y(\cdot)$ defined by $Y(t) := \tau - \bar{L}(t) - \theta_1 t/\alpha$, for $t \geq 0$. This gives $Z(t) = \tau + X(t) - \bar{L}(t) - \theta_1 t/\alpha = Q(t) - \theta_1 t/\alpha$. A similar argument as for the stopping time σ and (**??**) yield the system of equations, for each $i = 2, \ldots, n+1$.

$$e^{-\rho_k(\theta_2)(K-\tau)} h_{k,i}(\theta_2) - e^{-\rho_k(\theta_2)(K-\tau)} w_i(\theta_2) - \rho_k(\theta_2) \sum_{j=2}^{n+1} \bar{\ell}_j(\theta_2) h_{k,j}(\theta_2) = 0 \quad \text{for } k = 1, \ldots, n+1, \tag{3.8}$$

where $\bar{\ell}_j(\theta_2) = \mathbb{E}\left[\int_0^T e^{-\theta_2 s} 1_{\{\mathcal{J}(s)=j\}} \mathrm{d}\bar{L}(s)\right], j = 2, \ldots, n+1$. Using the method of determinants we can write $w_i(\theta_2)$ in the following form:

$$w_i(\theta_2) = \mathbb{E}\left[e^{-\theta_2 T} | \, \mathcal{J}(0) = i\right] = \frac{\sum_{k=1}^{n+1} (-1)^{1+k} c_k(\theta_2) e^{-\rho_k(\theta_2)(K-\tau)} h_{k,i}(\theta_2)}{\sum_{k=1}^{n+1} (-1)^{1+k} c_k(\theta_2) e^{-\rho_k(\theta_2)(K-\tau)}}, \tag{3.9}$$

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

$$c_k(\theta_2) = \begin{vmatrix} \rho_1(\theta_2) h_{1,2}(\theta_2) & \rho_1(\theta_2) h_{1,3}(\theta_2) & \ldots & \rho_1(\theta_2) h_{1,n+1}(\theta_2) \\ \rho_2(\theta_2) h_{2,2}(\theta_2) & \rho_2(\theta_2) h_{2,3}(\theta_2) & \ldots & \rho_2(\theta_2) h_{2,n+1}(\theta_2) \\ \vdots & \vdots & & \vdots \\ \rho_{k-1}(\theta_2) h_{k-1,2}(\theta_2) & \rho_{k-1}(\theta_2) h_{k-1,3}(\theta_2) & \ldots & \rho_{k-1}(\theta_2) h_{k-1,n+1}(\theta_2) \\ \rho_{k+1}(\theta_2) h_{j+1,2}(\theta_2) & \rho_{k+1}(\theta_2) h_{k+1,3}(\theta_2) & \ldots & \rho_{k+1}(\theta_2) h_{k+1,n+1}(\theta_2) \\ \vdots & \vdots & & \vdots \\ \rho_{n+1}(\theta_2) h_{n+1,2}(\theta_2) & \rho_{n+1}(\theta_2) h_{n+1,3}(\theta_2) & \ldots & \rho_{n+1}(\theta_2) h_{n+1,n+1}(\theta_2) \end{vmatrix} \tag{3.10}$$

Denoting

$$C(\theta_2) = \sum_{k=1}^{n+1} (-1)^{1+k} c_k(\theta_2) e^{-\rho_k(\theta_2)(K-\tau)} \tag{3.11}$$

and substituting the expression found for $w_i(\theta_2)$ in (**??**) into (**??**) yields the result of Theorem **??** with the $z_i(\theta_2)$, for $i = 2, \ldots, n+1$, given by the system of equations in (**??**). □

3.2 The finite buffer queue

Using the result of Theorem **??** 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 $(n, \vec{\alpha}_0, \mathbf{T})$. 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 $(n, \vec{\alpha}_0, \mathbf{T})$ distribution. Such a process has Laplace exponent equal to

$$\phi(\alpha) = -\alpha r_1 - \lambda + \lambda \hat{B}[\alpha] = -\alpha r_1 - \lambda + \lambda \vec{\alpha}_0^{\intercal} (\alpha \mathbf{I} - \mathbf{T})^{-1} \vec{t}, \tag{3.12}$$

where $\vec{t} = -\mathbf{T}\mathbf{1}$. As for the MAFP in Section **??** we determine the joint transform of the random variables U and D.

Corollary 3.1. *For $\theta_1, \theta_2 \geq 0$, the joint transform of the random variables D and U for a doubly reflected compound Poisson process is given by*

$$\mathbb{E}\, e^{-\theta_1 D - \theta_2 U} = \frac{1}{C(\theta_2)} \sum_{i=2}^{n+1} z_i(\theta_1) \sum_{j=1}^{n+1} (-1)^{j+1} c_j(\theta_2) e^{-p_j(\theta_2)(K-\tau)} h_{j,i}(\theta_2),$$

where $c_j(\theta_2), j = 1, \ldots, n+1$ and $C(\theta_2)$ are as in (??) and (??); $z_i(\theta_1)$ for $i = 2, \ldots, n+1$ are determined as the solution of a system of linear equations; this system is given in (??). The difference is that the roots $\rho_j(\theta_i)$, $i = 1, 2$, are replaced by the roots of the equation $\phi(\alpha) = \theta_i$, denoted by $p_j(\theta_i)$, $i = 1, 2$. The vectors $\vec{h}_k(\cdot)$ for $k = 1, \ldots, n+1$ are defined as follows:

$$h_{k,1}(\cdot) = 1 \;\; \forall k = 1, \ldots, n+1 \;\; \textit{and} \;\; h_{k,j}(\cdot) = \hat{B}_j[p_k(\cdot)] \;\; j = 2, \ldots, n+1.$$

The result in Corollary **??** can be derived as a limiting case of Theorem **??** by considering a suitable MAFP. The workload process $\{Q(t)\}_{t\geq 0}$ can be studied as the limit of a MAFP in the following sense. Following the construction presented in Section **??** 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

$$\mathcal{Q}^r = \begin{bmatrix} -\lambda & \lambda \vec{\alpha}_0^{\intercal} \\ r\vec{t} & r\mathbf{T} \end{bmatrix},$$

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 $\lambda, \vec{t}, \vec{\alpha}_0$ and $\mathbf{T}$ are the same as in Section **??**. 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 $r > 0$. Such a MAFP decreases linearly with rate r_1 during OFF-times, which are exponentially distributed with parameter λ and increases linearly with rate r during ON-times, which have a phase-type $(n, \vec{\alpha}_0, r\mathbf{T})$ distribution. By multiplying the matrix $\mathbf{T}$ with the factor r we see that the resulting phase-type distribution behaves like a phase-type distribution with representation $(n, \vec{\alpha}_0, \mathbf{T})$ divided by r. Using the representation in (**??**) we see that 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 $(n, \vec{\alpha}_0, \mathbf{T})$ 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 **??** and letting $r \to \infty$.

4 Discussion

In this note we have studied the occupation time of the set $[0, \tau]$ upto time t for the finite buffer fluid queue with phase type ON-times and the M/G/1 queue with phase-type jumps. Essential in our analysis was the joint transform of the consecutive periods below and above τ, i.e., $\mathbb{E}\, e^{-\theta_1 D - \theta_2 U}$ for $\theta_1 \geq 0, \theta_2 \geq 0$. The double transform of the occupation time uniquely specifies its distribution, which can be evaluated by numerically inverting the double transform [**?**]. Such a procedure has been carried out in [**?**] for the M/M/s queue, where an explicit expression for the double transform can be derived. For the current model, the transform is given implicitly, where for given θ_1, θ_2 linear equations need to be solved. An interesting topic for further research is

to evaluate how sensitive the numerical inversion algorithms are when the double transform is given implicitly. This issue also emerges in [?] where the author studies first hitting times for Lévy processes with phase-type jumps by deriving expressions for their Laplace transform.

Acknowledgements

We would like to thank the associate editor for his inspiring comments.

The research of N. Starreveld and M. Mandjes is partly funded by the NWO Gravitation project Networks, grant number 024.002.003.

References

[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 (200). *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] L. Breuer (2012). *Occupation Times for Markov-Modulated Brownian Motion*. Journal of Applied Probability, 49(2), pp. 549â€“565.

[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] 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.

[9] 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.

[10] K. Debicki and M. Mandjes (2015). *Queues and Lévy Fluctuation Theory*. Springer, New York.

[11] 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. 1776â€“1794.

[12] O. Kella (2006). *Reflecting thoughts*. Statistics and Probability Letters, 76(16), pp. 1808-1811.

[13] 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.

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

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

[16] 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.

[17] N. J. Starreveld, R. Bekker and M. Mandjes (2016). *Occupation times for regenerative processes with Lévy applications*. Submitted. arXiv:1602.05131.

[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.