J. Appl. Probab. 1–28 (2024)
*doi:*10.1017/jpr.2024.77

A QUEUE WITH INDEPENDENT AND IDENTICALLY DISTRIBUTED ARRIVALS

MICHEL MANDJES,[*] [**] *Leiden University and University of Amsterdam*
DANIËL T. RUTGERS,[*] [***] *Leiden University*

Abstract

In this paper we consider the workload of a storage system with the unconventional feature that the arrival times, rather than the interarrival times, are independent and identically distributed samples from a given distribution. We start by analyzing the 'base model' in which the arrival times are exponentially distributed, leading to a closed-form characterization of the queue's workload at a given moment in time (i.e. in terms of Laplace–Stieltjes transforms), assuming the initial workload was 0. Then we consider four more general models, each of them having a specific additional feature: (a) the initial workload being allowed to have any arbitrary non-negative value, (b) an additional stream of Poisson arrivals, (c) phase-type arrival times, (d) balking customers. For all four variants the transform of the transient workload is identified in closed form.

Keywords: Queueing; service systems; independent arrivals; workload; Laplace transforms

2020 Mathematics Subject Classification: Primary 60K25
Secondary 90B22; 44A10

1. Introduction

The model that we analyze in this paper is described as follows. We consider a queue in which there are $m \in \mathbb{N}$ arrivals, corresponding to independent and identically distributed (i.i.d.) arrival times that are sampled from a given distribution on the positive half-line; henceforth we let A denote a non-negative random variable distributed as a generic arrival time. The customers' service times are i.i.d. as well (and in addition independent of the arrival times), distributed as the non-negative random variable B. With the queue starting empty at time 0, our main objective is to evaluate the resulting queue's workload distribution, at any given point in time.

Note that when considering this model, we depart from the classical queueing paradigm in which the *interarrival* times are assumed to be i.i.d., rather than the arrival times. This model with i.i.d. interarrival times (or, equivalently, with renewal arrivals) is the intensively studied GI/G/1 queue. There are various compelling reasons to consider our model with i.i.d. arrival times. In the first place, it is observed that renewal arrivals are conceptually problematic, as they require any newly arriving customer to have knowledge of the arrival epoch of the previous

Received 23 January 2024; accepted 1 July 2024.

[*] Postal address: Mathematical Institute, Leiden University, PO Box 9512, 2300 RA Leiden, The Netherlands.
[**] Postal address: Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, The Netherlands. Email address: m.r.h.mandjes@math.leidenuniv.nl
[***] Email address: d.t.rutgers@math.leidenuniv.nl

customer (except in the case of Poisson arrivals, due to the memoryless property). In the second place, our model could be used to study the situation in which customers decide independently of each other when they want to use a given service. An example could relate to a scenario in which clients choose independently of each other when to visit a shop; the distribution of the arrival time A could reflect the day profile.

We proceed by providing a brief account of the existing literature. Despite the fact that the above model provides a highly natural description of a broad range of service systems, it is considerably less well understood than the more conventional class of queues with renewal-type arrivals. As the ith interarrival time is $A^{(i)} - A^{(i-1)} =: \Delta^{(i)}$, with $A^{(i)}$ denoting the ith-order statistic of the m arrival times, Honnappa *et al.*, in their influential paper [11], call the system an $\Delta^{(i)}$/G/1 queue. Other names have been used as well: in the terminology of the seminal paper by Glazer and Hassin [7] one would call the system a ?/G/1 queue, while Honnappa [10] later used RS/G/1 (RS standing for 'randomly scattered').

We do not provide an exhaustive overview of the results available, but restrict ourselves to a few recent key references; a more detailed overview can be found in [9, Section 2.1]. So far, hardly any explicit results are available for the transient queue length (i.e. the number of customers present at a given time t). Under an appropriately chosen scaling of the service-time distribution, [11] succeeds in developing fluid and diffusion approximations for the limiting regime as $m \to \infty$. In [3, 4], it is shown that the queueing process converges in a specific heavy-traffic regime to a reflected Brownian motion with non-linear drift. Sample-path large-deviation results have been established in [10]. Under the assumption that the service times are exponentially distributed, a system of Kolmogorov backward equations can be set up, so as to describe the transient queue-length distribution; this method is due to [7] and was further generalized in [14]. As described in great detail in [9], there is a strong relation to the strategic queueing game in which each customer has to decide when to arrive, without any coordination with the other customers.

We conclude this introduction by detailing this paper's contributions and organization. As stated in [11], 'exact analysis of this model is impossible for general service processes'. While this remains true, to date, for the transient *queue-length* distribution, the results of this paper show that one can provide a full analysis of the transient *workload* distribution, albeit in terms of transforms. Specifically, for the case of exponentially distributed arrival times, in Section 2 we develop a technique that provides the Laplace–Stieltjes transform of the workload at an exponentially distributed point in time. As we demonstrate, relying on the powerful computational techniques of [1] and [12], this enables us to numerically evaluate various relevant workload-related performance metrics.

The second contribution concerns various extensions of this 'base model'. (a) In the first place we consider in Section 3 the model in which the workload is not necessarily starting empty at time 0. The analysis relies on relations between the workload process and the associated non-reflected process and on a description of the corresponding first-passage time process as a Markov additive process [5]. (b) Second, in Section 4 we allow an additional Poisson arrival stream, with i.i.d. service times (not necessarily distributed as the service time B of the finite pool of m customers). The analysis reduces to solving a recursion involving m unknown constants that can be identified using a result from [13]. (c) Then, in Section 5, we allow the arrival times to be of phase type; this class of distributions is particularly relevant as any random variable on the positive half-line can be approximated arbitrarily closely by a phase-type random variable [2, Section III.4]. Also, in the analysis of this case with phase-type arrivals, unknown constants appear, which can again be determined relying on [13]. (d) The last variant

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

we consider, in Section 6, is the one where, based on the workload they face when arriving, customers decide whether or not to enter the system. This concept is often referred to as *balking*, and has been studied in various settings; see e.g. the seminal work [8].

2. Exponentially distributed arrival times

This section focuses on the 'base model', in which the generic arrival time A has an exponential distribution with parameter $\lambda > 0$. Our objective is to uniquely characterize the queue's transient workload. We do so by identifying a closed-form expression for the Laplace–Stieltjes transform of the queue's workload after an exponentially distributed time interval. Throughout, it is assumed that the system starts empty at time 0.

Let the cumulative distribution function of the service times be $B(\cdot)$. Our focus lies on finding a probabilistic characterization of the workload process $(W(t))_{t\geq 0}$. With $N(t) \in \{0, \ldots, m\}$ denoting the number of clients that have *not* arrived by time $t \geq 0$, the key object of study is the cumulative distribution function

$$F_t(x, n) := \mathbb{P}(W(t) \leq x, N(t) = n).$$

It is noted that $W(t)$ has an atom in 0, in that $\mathbb{P}(W(t) = 0) > 0$ for all $t > 0$. For $x > 0$, $t > 0$, and $n \in \{0, \ldots, m\}$, we introduce the corresponding density

$$f_t(x, n) := \frac{\partial}{\partial x}\mathbb{P}(W(t) \leq x, N(t) = n) = \frac{\partial}{\partial x}F_t(x, n),$$

and for $t > 0$ and $n \in \{0, \ldots, m\}$ the zero-workload probabilities

$$P_t(n) := \mathbb{P}(W(t) = 0, N(t) = n).$$

Remark 2.1. When considering exponentially distributed arrival times, the overall arrival process considered in our model is equivalent to a non-homogeneous Poisson arrival process, conditioned on m arrivals in total, with arrival rate function $\lambda(t) = C\mathrm{e}^{-\lambda t}$ (where the constant $C > 0$ is arbitrary). This observation was made before in [17].

2.1. Setting up the differential equation

The distribution function $F_t(x, n)$ can be analyzed by a classical approach in which its value at time $t + \Delta t$ is related to its counterpart at time t. Indeed, observe that, as $\Delta t \downarrow 0$, for any $x > 0$, $t > 0$, and $n \in \{0, \ldots, m-1\}$,

$$\begin{aligned} F_{t+\Delta t}(x, n) = {} & (1 - \lambda n\, \Delta t) \cdot F_t(x + \Delta t, n) + \lambda(n+1)\, \Delta t \int_{(0,x]} f_t(y, n+1)\, B(x-y)\, \mathrm{d}y \\ & + \lambda(n+1)\, \Delta t\, P_t(n+1)\, B(x) + \mathrm{o}(\Delta t). \end{aligned} \tag{2.1}$$

Equation (2.1) has the following straightforward interpretation: the first term on the right-hand side represents the scenario of no arrival between t and $t + \Delta t$, the second term the scenario with one arrival in combination with the workload prior to the arrival being positive, and the third term the scenario with one arrival in combination with the workload prior to the arrival being zero; we remark that scenarios with more than one arrival are absorbed in the $\mathrm{o}(\Delta t)$ term.

After subtracting $F_t(x, n)$ from both sides of (2.1), dividing the full equation by Δt, and sending Δt to 0, we obtain the following partial differential equation: for any $x > 0$, $t > 0$, and

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

$n \in \{0, \ldots, m-1\}$,

$$\frac{\partial}{\partial t} F_t(x,n) - f_t(x,n) = -\lambda n\, F_t(x,n) + \lambda(n+1) \int_{(0,x]} f_t(y, n+1)\, B(x-y)\, \mathrm{d}y + \lambda(n+1)\, P_t(n+1)\, B(x). \tag{2.2}$$

2.2. Double transform

In order to uniquely characterize the solution of this partial differential equation, we work with a double transform. To this end, we first multiply the full equation (2.2) by $\mathrm{e}^{-\alpha x}$, for $\alpha \geq 0$, and integrate over $x \in (0, \infty)$, so as to convert the partial differential equation into an ordinary differential equation. In this analysis, we intensively work with the object

$$\mathscr{F}_t(\alpha, n) := \int_{(0,\infty)} \mathrm{e}^{-\alpha x} f_t(x,n)\, \mathrm{d}x.$$

By applying integration by parts, it is readily verified that, for any $n \in \{0, \ldots, m-1\}$,

$$\int_{(0,\infty)} \mathrm{e}^{-\alpha x}\, F_t(x,n)\, \mathrm{d}x = \frac{P_t(n) + \mathscr{F}_t(\alpha, n)}{\alpha}.$$

By applying Fubini, and denoting $\mathscr{B}(\alpha) := \mathbb{E}\, \mathrm{e}^{-\alpha B}$, again for any $n \in \{0, \ldots, m-1\}$,

$$\int_{(0,\infty)} \mathrm{e}^{-\alpha x} \int_{(0,x]} f_t(y,n)\, B(x-y)\, \mathrm{d}y\, \mathrm{d}x = \frac{\mathscr{F}_t(\alpha, n)\, \mathscr{B}(\alpha)}{\alpha}.$$

Upon combining the above identities, we readily arrive at the following (ordinary) differential equation:

$$\frac{\partial}{\partial t} \frac{P_t(n) + \mathscr{F}_t(\alpha, n)}{\alpha} - \mathscr{F}_t(\alpha, n) = -\lambda n\, \frac{P_t(n) + \mathscr{F}_t(\alpha, n)}{\alpha} + \lambda(n+1)\, \frac{(P_t(n+1) + \mathscr{F}_t(\alpha, n+1))\, \mathscr{B}(\alpha)}{\alpha},$$

which, with $\bar{\mathscr{F}}_t(\alpha, n) := P_t(n) + \mathscr{F}_t(\alpha, n)$, simplifies to

$$\frac{\partial}{\partial t} \frac{\bar{\mathscr{F}}_t(\alpha, n)}{\alpha} - \bar{\mathscr{F}}_t(\alpha, n) + P_t(n) = -\lambda n \frac{\bar{\mathscr{F}}_t(\alpha, n)}{\alpha} + \lambda(n+1)\, \frac{\bar{\mathscr{F}}_t(\alpha, n+1)\, \mathscr{B}(\alpha)}{\alpha},$$

The next step is to transform once more: we multiply the full equation in the previous display by $\mathrm{e}^{-\beta t}$, for $\beta > 0$, and integrate over $t \in (0, \infty)$, with the objective of turning the ordinary differential equation of the previous display into an algebraic equation. We use the notation

$$\mathscr{P}_n(\beta) := \int_{(0,\infty)} \mathrm{e}^{-\beta t}\, P_t(n)\, \mathrm{d}t,$$

$$\mathscr{G}_n(\alpha, \beta) \equiv \mathscr{G}_n(\alpha, \beta \mid m) := \int_{(0,\infty)} \mathrm{e}^{-\beta t}\, \bar{\mathscr{F}}_t(\alpha, n)\, \mathrm{d}t.$$

Using the same techniques as above, for $n \in \{0, \ldots, m-1\}$,

$$(\beta - \alpha)\mathscr{G}_n(\alpha, \beta) + \alpha\, \mathscr{P}_n(\beta) = -\lambda n\, \mathscr{G}_n(\alpha, \beta) + \lambda(n+1)\, \mathscr{G}_{n+1}(\alpha, \beta)\, \mathscr{B}(\alpha), \tag{2.3}$$

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

so that we arrive at the recursion

$$\mathscr{G}_n(\alpha,\beta)=\frac{\lambda(n+1)\,\mathscr{G}_{n+1}(\alpha,\beta)\,\mathscr{B}(\alpha)-\alpha\,\mathscr{P}_n(\beta)}{\beta-\alpha+\lambda n}. \tag{2.4}$$

The case $n=m$ can be dealt with explicitly. Indeed, observing that if $N(t)=m$ no arrival can have occurred before time t, we find

$$\mathscr{G}_m(\alpha,\beta)=\mathscr{P}_m(\beta)=\frac{1}{\beta+\lambda m}.$$

Remark 2.2. Note that the object $\beta\,\mathscr{G}_n(\alpha,\beta\mid m)$ has an appealing interpretation:

$$\beta\,\mathscr{G}_n(\alpha,\beta\mid m)=\int_{(0,\infty)}\beta\mathrm{e}^{-\beta t}\,\bar{\mathscr{F}}_t(\alpha,n)\,\mathrm{d}t=\mathbb{E}\big(\mathrm{e}^{-\alpha W(T_\beta)}\,\mathbf{1}_{\{N(T_\beta)=n\}}\big),$$

with T_β an exponentially distributed random variable with mean β^{-1}, independent of anything else. This observation will be intensively relied upon in Section 3.

2.3. Explicit solution

The recursion (2.4) can be readily solved, by repeated insertion. The eventual result is given in Theorem 2.1, but we first sketch the underlying approach, to explicitly identify all expressions involved.

Denoting the coefficients in the recursion by

$$\gamma_n\equiv\gamma_n(\alpha,\beta):=\frac{\lambda(n+1)\,\mathscr{B}(\alpha)}{\beta-\alpha+\lambda n},\qquad \delta_n\equiv\delta_n(\alpha,\beta):=-\frac{\alpha\,\mathscr{P}_n(\beta)}{\beta-\alpha+\lambda n},$$

we obtain the standard solution

$$\mathscr{G}_n(\alpha,\beta)=\mathscr{G}_m(\alpha,\beta)\prod_{i=n}^{m-1}\gamma_i+\sum_{j=n}^{m-1}\delta_j\prod_{i=n}^{j-1}\gamma_i, \tag{2.5}$$

following the convention that the empty product is defined as one. At this point, it is left to determine the unknown functions $\mathscr{P}_n(\beta)$, for $n=0,\dots,m-1$. That can be done by noting that any root of the denominator should be a root of the numerator too. To this end, observe that we can rewrite the expression for $\mathscr{G}_0(\alpha,\beta)$ in the form

$$\mathscr{G}_0(\alpha,\beta)=\frac{\mathscr{H}_m(\alpha,\beta)+\sum_{n=0}^{m-1}\mathscr{H}_n(\alpha,\beta)\,\mathscr{P}_n(\beta)}{\prod_{n=0}^{m-1}(\beta-\alpha+\lambda n)}, \tag{2.6}$$

for appropriately chosen functions $\mathscr{H}_n(\alpha,\beta)$, with $n=0,1,\dots,m$. Note that the (distinct) roots of the denominator are $\alpha_j:=\beta+\lambda j>0$, with $j=0,\dots,m-1$. This means that the unknown functions $\mathscr{P}_n(\beta)$ can be found by solving the following linear equations: for $j=0,\dots,m-1$,

$$-\mathscr{H}_m(\alpha_j,\beta)=\sum_{n=0}^{m-1}\mathscr{H}_n(\alpha_j,\beta)\,\mathscr{P}_n(\beta). \tag{2.7}$$

Hence we can find the m unknowns from these m equations.

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

We proceed by explicitly identifying the above objects. In these derivations, we intensively use the compact notations

$$\xi_n \equiv \xi_n(\alpha, \beta) := \prod_{i=0}^{n} (\beta - \alpha + \lambda i), \qquad \eta_n \equiv \eta_n(\alpha, \beta) := \prod_{i=0}^{n} (\lambda(i+1)\mathscr{B}(\alpha)),$$

with empty products being defined as 1. We can rewrite (2.5) in the form of (2.6), as follows:

$$\begin{aligned}\mathscr{G}_0(\alpha, \beta) &= \mathscr{G}_m(\alpha, \beta)\frac{\prod_{n=0}^{m-1} \lambda(n+1)\mathscr{B}(\alpha)}{\prod_{n=0}^{m-1} (\beta - \alpha + \lambda n)} + \sum_{j=0}^{m-1} \delta_j \frac{\prod_{n=0}^{j-1} \lambda(n+1)\mathscr{B}(\alpha)}{\prod_{n=0}^{j-1} (\beta - \alpha + \lambda n)} \\ &= \frac{\mathscr{H}_m(\alpha, \beta) + \sum_{j=0}^{m-1} \delta_j(\xi_{m-1}/\xi_{j-1})\eta_{j-1}}{\xi_{m-1}} \\ &= \frac{\mathscr{H}_m(\alpha, \beta) - \sum_{j=0}^{m-1} \alpha(\xi_{m-1}/\xi_j)\eta_{j-1}\ \mathscr{P}_j(\beta)}{\xi_{m-1}} \\ &= \frac{\mathscr{H}_m(\alpha, \beta) + \sum_{n=0}^{m-1} \mathscr{H}_n(\alpha, \beta)\mathscr{P}_n(\beta)}{\xi_{m-1}},\end{aligned}$$

where

$$\mathscr{H}_m(\alpha, \beta) := \mathscr{G}_m(\alpha, \beta)\, \eta_{m-1} = \mathscr{G}_m(\alpha, \beta)(\lambda\mathscr{B}(\alpha))^m m!$$

and, for $n \in \{0, \ldots, m-1\}$,

$$\mathscr{H}_n(\alpha, \beta) := -\alpha\left(\prod_{i=n+1}^{m-1} (\beta - \alpha + \lambda i)\right)\left(\prod_{i=0}^{n-1} (\lambda(i+1)\mathscr{B}(\alpha))\right) = -\alpha(\lambda\mathscr{B}(\alpha))^n n!\, \frac{\xi_{m-1}}{\xi_n}.$$

Given these expressions for $\mathscr{H}_n(\alpha, \beta)$, we can now identify expressions for $\mathscr{P}_n(\beta)$ as well. To this end, note that $\mathscr{H}_n(\alpha_j, \beta) = 0$ for $j \in \{n+1, \ldots, m-1\}$, because α_j can be a root of the first product in the definition of $\mathscr{H}_n(\alpha_j, \beta)$.

We first determine $\mathscr{P}_{m-1}(\beta)$. Substituting α_{m-1} into (2.7) gives

$$-\mathscr{H}_m(\alpha_{m-1}, \beta) = \sum_{n=0}^{m-1} \mathscr{H}_n(\alpha_{m-1}, \beta)\ \mathscr{P}_n(\beta) = \mathscr{H}_{m-1}(\alpha_{m-1}, \beta)\ \mathscr{P}_{m-1}(\beta),$$

so that

$$\mathscr{P}_{m-1}(\beta) = -\frac{\mathscr{H}_m(\alpha_{m-1}, \beta)}{\mathscr{H}_{m-1}(\alpha_{m-1}, \beta)}.$$

For $n \in \{0, \ldots, m-1\}$ we then have

$$\mathscr{P}_n(\beta) = -\frac{\mathscr{H}_m(\alpha_n, \beta) + \sum_{i=n+1}^{m-1} \mathscr{H}_i(\alpha_n, \beta)\mathscr{P}_i(\beta)}{\mathscr{H}_n(\alpha_n, \beta)}. \tag{2.8}$$

This means that the linear system (2.7) can be solved recursively: when plugging $n = m-2$ into equation (2.8), we obtain $\mathscr{P}_{m-2}(\beta)$ in terms of $\mathscr{P}_{m-1}(\beta)$, after which $\mathscr{P}_{m-3}(\beta)$ can be

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

expressed in terms of $\mathscr{P}_{m-2}(\beta)$ and $\mathscr{P}_{m-1}(\beta)$, and so on. The next theorem summarizes our findings thus far.

Theorem 2.1. (Base model.) *For any $\alpha \geq 0$ and $\beta > 0$, and $n \in \{0, \ldots, m-1\}$, the transform $\mathscr{G}_n(\alpha, \beta \mid m)$ is given by* (2.5), *where the transforms $\mathscr{P}_0(\beta), \ldots, \mathscr{P}_{m-1}(\beta)$ follow from the recursion* (2.8), *and $\mathscr{G}_m(\alpha, \beta \mid m) = \mathscr{P}_m(\beta) = (\beta + \lambda m)^{-1}$.*

Remark 2.3. There is a related way to derive, for $n \in \{0, \ldots, m-1\}$, expressions for the transforms $\mathscr{P}_n(\beta)$. To this end, note that for the root of the denominator in (2.4), i.e. $\alpha_n = \beta + \lambda n$, the numerator must also equal zero. This leads to the relation

$$\lambda(n+1)\, \mathscr{G}_{n+1}(\alpha_n, \beta)\mathscr{B}(\alpha_n) - \alpha_n \mathscr{P}_n(\beta) = 0,$$

which rewritten gives

$$\mathscr{P}_n(\beta) = \lambda(n+1)\frac{\mathscr{G}_{n+1}(\alpha_n, \beta)\mathscr{B}(\alpha_n)}{\alpha_n}. \tag{2.9}$$

Together with equation (2.4) and $\mathscr{G}_m(\alpha, \beta) = \mathscr{P}_m(\beta) = (\beta + \lambda m)^{-1}$, (2.9) can be solved recursively as well. Specifically, equations (2.4) and (2.9) are to be applied alternately: from the known expression for $\mathscr{G}_m(\alpha, \beta)$ we find $\mathscr{P}_{m-1}(\beta)$ by (2.9), then $\mathscr{G}_{m-1}(\alpha, \beta)$ (for any $\alpha \geq 0$) follows from (2.4), then $\mathscr{P}_{m-2}(\beta)$ again by (2.9), and so on.

2.4. Alternative approach

We now detail an alternative procedure by which the transforms $\mathscr{P}_0(\beta), \ldots, \mathscr{P}_{m-1}(\beta)$ can be determined. The main reason why we include it here is that in Section 4 we will intensively rely on the underlying argumentation; the account below serves to introduce the concepts in an elementary setting.

Denote $\boldsymbol{V}(\alpha, \beta) = (\mathscr{V}_0(\alpha, \beta), \ldots, \mathscr{V}_{m-1}(\alpha, \beta))^\top$, where the ith component is given by

$$\mathscr{V}_i(\alpha, \beta) = \alpha \mathscr{P}_i(\beta) - \frac{\lambda m}{\beta + \lambda m}\mathscr{B}(\alpha)\mathbf{1}_{\{i=m-1\}}.$$

In addition, $\boldsymbol{G}(\alpha, \beta) = (\mathscr{G}_0(\alpha, \beta), \ldots, \mathscr{G}_{m-1}(\alpha, \beta))^\top$. Then it is easily checked that the system of equations (2.3) can be rewritten in matrix–vector notation as

$$M(\alpha, \beta)\, \boldsymbol{G}(\alpha, \beta) = \boldsymbol{V}(\alpha, \beta).$$

Here the $(i, i+1)$th entry (for $i \in \{0, \ldots, m-2\}$) of the $m \times m$ matrix

$$M(\alpha, \beta) = (M_{ij}(\alpha, \beta))_{i,j=0}^{m-1}$$

is given by $\lambda(i+1)\, \mathscr{B}(\alpha)$, and the (i, i)th entry (for $i \in \{0, \ldots, m-1\}$) by $\alpha - \beta - \lambda i$.

The next observation is that the transpose of $M(\alpha, \beta)$ is the so-called matrix exponent of a Markov additive process (MAP) [2, Section XI.2]. This can be seen as follows.

- A MAP is defined by a dimension $d \in \mathbb{N}$, a possibly defective $d \times d$ transition rate matrix Q governing a background process, jump sizes J_{ij} corresponding to transitions by the background process from state i to state j (with $i, j \in \{1, \ldots, d\}$ with $i \neq j$), and Lévy processes $Y_i(t)$ with Laplace exponents $\varphi_i(\cdot)$ that are active when the background process is in state i (with $i \in \{1, \ldots, d\}$). When the jumps are non-negative and the Lévy processes spectrally positive, and when imposing killing at rate $\beta > 0$, the matrix exponent

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

of this MAP, for $\alpha \geq 0$, is given by

$$\text{diag}\{\varphi_1(\alpha) - \beta, \dots, \varphi_d(\alpha) - \beta\} + Q \circ \mathscr{J}(\alpha), \tag{2.10}$$

with $\circ$ denoting the Hadamard product, and the (i, j)th entry of $\mathscr{J}(\alpha)$ defined by $\mathbb{E}\,\mathrm{e}^{-\alpha J_{ij}}$.

- Then one can directly verify that the transpose of our matrix $M(\alpha, \beta)$ is of the form (2.10); recall that α is the Laplace exponent of a deterministic drift of rate 1.

We proceed by studying the roots of $\det M(\alpha, \beta)$ (for any given $\beta > 0$), which evidently coincide with the roots of $\det M(\alpha, \beta)^\top$. Applying the machinery developed in [13] and [15] for matrix exponents of Markov additive processes, we conclude that it has m roots in the right-half of the complex α-plane, say $\alpha_0, \dots, \alpha_{m-1}$. Technically, our instance fits into the framework of [15, Proposition 2], in that the underlying background process is not irreducible (with state 0 being an absorbing state).

The next step is to observe that, by Cramer's rule, for $n \in \{0, \dots, m-1\}$,

$$\mathscr{G}_n(\alpha, \beta) = \frac{\det M_n(\alpha, \beta)}{\det M(\alpha, \beta)},$$

where $M_n(\alpha, \beta)$ is defined as $M(\alpha, \beta)$, with the nth column replaced by $\boldsymbol{V}(\alpha, \beta)$. This means that, for $j \in \{0, \dots, m-1\}$, $\det M_n(\alpha_j, \beta) = 0$ for $n \in \{0, \dots, m-1\}$. This seemingly leads to m^2 (linear) equations in the m unknowns $\mathscr{P}_0(\beta), \dots, \mathscr{P}_{m-1}(\beta)$, but it turns out that all equations corresponding to the same α_j effectively contain the same information: if $\det M(\alpha, \beta) = \det M_n(\alpha, \beta) = 0$, then also $\det M_{n'}(\alpha, \beta) = 0$ for $n' \neq n$; to see this, exactly the same reasoning as in [15, Section 3.3.1] can be followed.

In our specific case the roots $\alpha_j = \beta + \lambda j$, for $j \in \{0, \dots, m-1\}$, are distinct. We thus end up with m linear equations in equally many unknowns. We conclude this section by applying the above method, so as to recover our previous result (2.8).

Lemma 2.1. *The determinant of the matrix $M_n(\alpha, \beta)$ defined above is given by, for $n \in \{0, \dots, m-1\}$,*

$$\det M_n(\alpha, \beta) = C_n(\alpha, \beta) \prod_{i=0,\, i\neq n}^{m-1} (\alpha - \beta - \lambda i),$$

with

$$C_n(\alpha, \beta) := \alpha\left(\mathscr{P}_n(\beta) + \frac{\mathscr{H}_m(\alpha, \beta) + \sum_{i=n+1}^{m-1} \mathscr{H}_i(\alpha, \beta)\mathscr{P}_i(\beta)}{\mathscr{H}_n(\alpha, \beta)}\right).$$

Proof. See Appendix A. □

According to the above recipe, for all $j \in \{0, \dots, m-1\}$ we necessarily have $\det M_n(\alpha_j, \beta) = 0$. From Lemma 2.1 we find that this is indeed the case for all $\alpha_j = \beta + \lambda j$ with $j \neq n$, as they are the roots of the product term appearing in $\det M_n(\alpha, \beta)$. Inserting $j = n$ into $\det M_n(\alpha_j, \beta) = 0$, we conclude that $C_n(\alpha_n, \beta) = 0$, from which it follows that (2.8) applies.

In Figure 1 we plot, for different values of the number of customers m, the mean workload $\mathbb{E}\, W(t)$, its variance $\text{Var}\, W(t)$, and the probability of an empty buffer $\mathbb{P}(W(t) = 0)$, as functions of time. These are numerically obtained by converting, in the obvious manner, the recursion

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

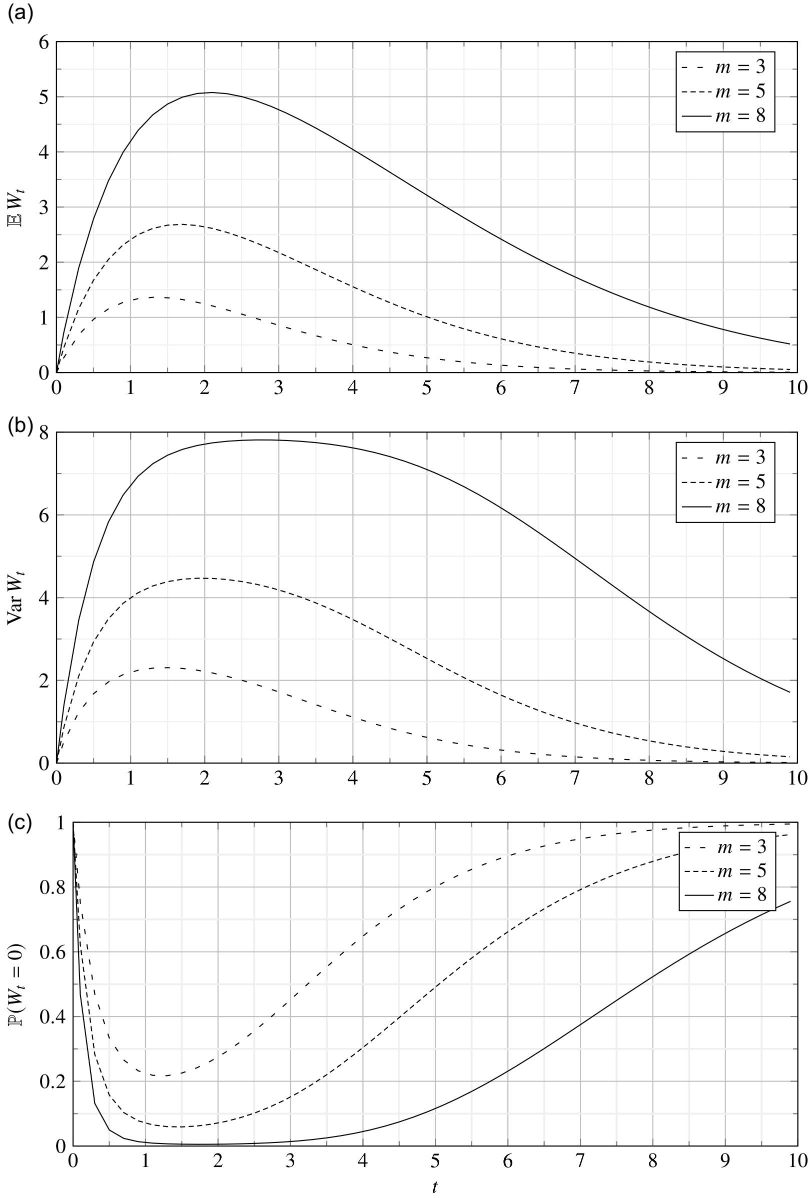

FIGURE 1. Mean workload (a), variance of the workload (b), and empty-buffer probability (c) in the base model, as functions of time, for different values of m.

for the double transform into recursions involving

$$\int_0^\infty \mathrm{e}^{-\beta t}\mathbb{E}\big(W(t)\,\mathbf{1}_{\{N(t)=n\}}\big)\,\mathrm{d}t, \qquad \int_0^\infty \mathrm{e}^{-\beta t}\mathbb{E}\big(W(t)^2\,\mathbf{1}_{\{N(t)=n\}}\big)\,\mathrm{d}t$$

$$\text{and} \quad \int_0^\infty \mathrm{e}^{-\beta t}\mathbb{P}(W(t)=0, N(t)=n)\,\mathrm{d}t,$$

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

and then perform numerical Laplace inversion [1] with respect to β. The service times are exponentially distributed, and we have used $\lambda = \mu = 1$.

3. Starting at arbitrary initial workload

So far we have assumed that at time zero the workload level equals zero. In this section we generalize this to cover the case in which we start at any initial workload level $x \geq 0$. The object of our interest is

$$\bar{\mathscr{G}}_n(\alpha,\, \beta \mid x,\, m) := \mathbb{E}\big(\mathrm{e}^{-\alpha W(T_\beta)}\mathbf{1}_{\{N(T_\beta)=n\}} \mid W(0) = x,\, N(0) = m\big); \tag{3.1}$$

here T_β again denotes an exponentially distributed random variable with mean β^{-1}, independent of anything else. Henceforth we alternatively denote the right-hand side of (3.1) by

$$\mathbb{E}_{x,m}\big(\mathrm{e}^{-\alpha W(T_\beta)}\mathbf{1}_{\{N(T_\beta)=n\}}\big),$$

that is, we systematically use subscripts to indicate the initial conditions.

3.1. Derivation of the transform

The workload process $W(t)$ is often referred to as the *reflected process*. The process cannot drop below zero; when there is no work in the system and there is service capacity available, the workload level remains 0. In the present section, we intensively work with the process $Y(\cdot)$, to be interpreted as the position of the associated *free process* (or non-reflected process). In particular, this means that, for any given $t \geq 0$, $Y(t)$ represents the the amount of work that has arrived in $(0,t]$ minus what could potentially have been served (i.e. t).

We further define the stopping time $\sigma(x) := \inf\{t\colon\ Y(t) < -x\}$, which is the first time that the buffer is empty given that the initial workload is x. Observe that $Y(\sigma(x)) = -x$ almost surely, as the process $Y(t)$ has no negative jumps.

We also work with the counterpart of (3.1) for the free process:

$$\check{\mathscr{G}}_n(\alpha,\, \beta \mid m) := \mathbb{E}\big(\mathrm{e}^{-\alpha Y(T_\beta)}\mathbf{1}_{\{N(T_\beta)=n\}} \mid N(0) = m\big).$$

So as to analyze the quantity under study, we distinguish between two disjoint scenarios: the scenario in which the workload has idled before T_β and its complement. This means that we split $\bar{\mathscr{G}}_n(\alpha,\, \beta \mid x,\, m) = \bar{\mathscr{G}}_n^-(\alpha,\, \beta \mid x,\, m) + \bar{\mathscr{G}}_n^+(\alpha,\, \beta \mid x,\, m)$, with, in self-evident notation,

$$\bar{\mathscr{G}}_n^-(\alpha,\, \beta \mid x,\, m) := \mathbb{E}_{x,m}\big(\mathrm{e}^{-\alpha W(T_\beta)}\mathbf{1}_{\{N(T_\beta)=n,\sigma(x)\leq T_\beta\}}\big),$$

$$\bar{\mathscr{G}}_n^+(\alpha,\, \beta \mid x,\, m) := \mathbb{E}_{x,m}\big(\mathrm{e}^{-\alpha W(T_\beta)}\mathbf{1}_{\{N(T_\beta)=n,\sigma(x)> T_\beta\}}\big).$$

We evaluate the objects $\bar{\mathscr{G}}_n^-(\alpha,\, \beta \mid x,\, m)$ and $\bar{\mathscr{G}}_n^+(\alpha,\, \beta \mid x,\, m)$ separately.

Observe that, by the strong Markov property in combination with the memoryless property of T_β and the arrival times,

$$\bar{\mathscr{G}}_n^-(\alpha,\, \beta \mid x,\, m) = \sum_{k=n}^{m} \mathbb{P}_m(N(\sigma(x)) = k,\, \sigma(x) \leq T_\beta)\, \bar{\mathscr{G}}_n(\alpha,\, \beta \mid 0,\, k), \tag{3.2}$$

where we already know from Theorem 2.1 how we can evaluate $\bar{\mathscr{G}}_n(\alpha, \beta \mid 0, k) = \beta\, \mathscr{G}_n(\alpha,\, \beta \mid k)$; see Remark 2.2. The interpretation underlying the decomposition on the right-hand side of (3.2) is that the queue starts empty at time $\sigma(x)$.

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

We proceed by analyzing the remaining quantity, $\bar{\mathscr{G}}_n^+(\alpha,\,\beta\mid x,\,m)$. To this end, we first observe that $\check{\mathscr{G}}_n(\alpha,\,\beta\mid m)=\check{\mathscr{G}}_n^-(\alpha,\,\beta\mid x,\,m)+\check{\mathscr{G}}_n^+(\alpha,\,\beta\mid x,\,m)$, where, in self-evident notation,

$$\check{\mathscr{G}}_n^-(\alpha,\,\beta\mid x,\,m):=\mathbb{E}_m\big(\mathrm{e}^{-\alpha Y(T_\beta)}\mathbf{1}_{\{N(T_\beta)=n,\sigma(x)\le T_\beta\}}\big),$$

$$\check{\mathscr{G}}_n^+(\alpha,\,\beta\mid x,\,m):=\mathbb{E}_m\big(\mathrm{e}^{-\alpha Y(T_\beta)}\mathbf{1}_{\{N(T_\beta)=n,\sigma(x)>T_\beta\}}\big).$$

The crucial step is that on the event $\{\sigma(x)>T_\beta\}$ it holds that $W(T_\beta)=x+Y(T_\beta)$. As a consequence,

$$\bar{\mathscr{G}}_n^+(\alpha,\,\beta\mid x,\,m)=\mathrm{e}^{-\alpha x}\,\check{\mathscr{G}}_n^+(\alpha,\,\beta\mid x,\,m)=\mathrm{e}^{-\alpha x}\,\check{\mathscr{G}}_n(\alpha,\,\beta\mid m)-\mathrm{e}^{-\alpha x}\,\check{\mathscr{G}}_n^-(\alpha,\,\beta\mid x,\,m).$$

The second term on the right-hand side of the previous display can be further evaluated, using $Y(\sigma(x))=-x$ in combination with the memoryless property. Using the same reasoning as before, we thus find

$$\check{\mathscr{G}}_n^-(\alpha,\,\beta\mid x,\,m)=\mathrm{e}^{\alpha x}\sum_{k=n}^{m}\mathbb{P}_m(N(\sigma(x))=k,\,\sigma(x)\le T_\beta)\,\check{\mathscr{G}}_n(\alpha,\,\beta\mid k).$$

From the above, we conclude that it suffices to be able to evaluate the object, for $k\in\{0,\,\ldots,\,m\}$ and $n\in\{0,\,\ldots,\,k\}$,

$$p_{m,k}(x,\,\beta):=\mathbb{P}_m(N(\sigma(x))=k,\,\sigma(x)\le T_\beta)\quad\text{and}\quad\check{\mathscr{G}}_n(\alpha,\,\beta\mid k).$$

The latter quantity can be evaluated by solving a system of linear equations, while the former is slightly harder to analyze.

- It is readily verified that, for $n\in\{0,\,\ldots,\,k\}$,

  $$\check{\mathscr{G}}_n(\alpha,\,\beta\mid k)=\frac{\lambda k}{\lambda k+\beta-\alpha}\mathscr{B}(\alpha)\,\check{\mathscr{G}}_n(\alpha,\,\beta\mid k-1)\,\mathbf{1}_{\{k>n\}}+\frac{\beta}{\lambda k+\beta-\alpha}\,\mathbf{1}_{\{k=n\}}.\tag{3.3}$$

  Writing this system in the usual matrix–vector form, it is readily seen that it is diagonally dominant, ensuring the system has a unique solution. Alternatively, one can solve the system recursively (starting at $k=n$).

- We apply results from [5] to identify the probabilities $p_{m,k}(x,\,\beta)$ for $k\in\{0,\,\ldots,\,m\}$. We first observe that $(\sigma(x),\,N(\sigma(x)))$ is a MAP in $x\ge 0$ [5, Section 1.2]; note that this MAP does not have any non-decreasing subordinator states.
  To identify its characteristics, we first define the MAP corresponding to the free process $Y(t)$. To this end, we introduce a matrix $K(\alpha,\,\beta)$ with the $(i,\,i)$th entry given by $\alpha-\lambda i-\beta$ (for $i\in\{0,\,\ldots,\,m\}$), and the $(i,\,i-1)$th entry given by $\lambda i\,\mathscr{B}(\alpha)$ (for $i\in\{1,\,\ldots,\,m\}$), and all other entries equal to 0. The roots, for a given value of $\beta>0$, of $\det K(\alpha,\,\beta)=0$ are $\boldsymbol{d}(\beta):=(\beta,\,\lambda+\beta,\,\ldots,\,\lambda m+\beta)^\top$. The corresponding eigenvectors, solving $K(\alpha)\,\boldsymbol{v}=0$, can be evaluated recursively; calling these $\boldsymbol{v}_0,\,\ldots,\,\boldsymbol{v}_m$, we let V be a matrix of which the columns are these vectors. Then, by [5, equation (2)], in combination with [5, Theorem 1] and the fact that $Y(t)$ does not have any non-decreasing subordinator states, we obtain

  $$p_{m,k}(x,\,\beta)=\big(\exp(-\,VD(\beta)\,V^{-1}\,x)\big)_{m,k}=\big(V\exp(-\,D(\beta)\,x)V^{-1}\big)_{m,k},\tag{3.4}$$

  with $D(\beta):=\operatorname{diag}\{\boldsymbol{d}(\beta)\}$.

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

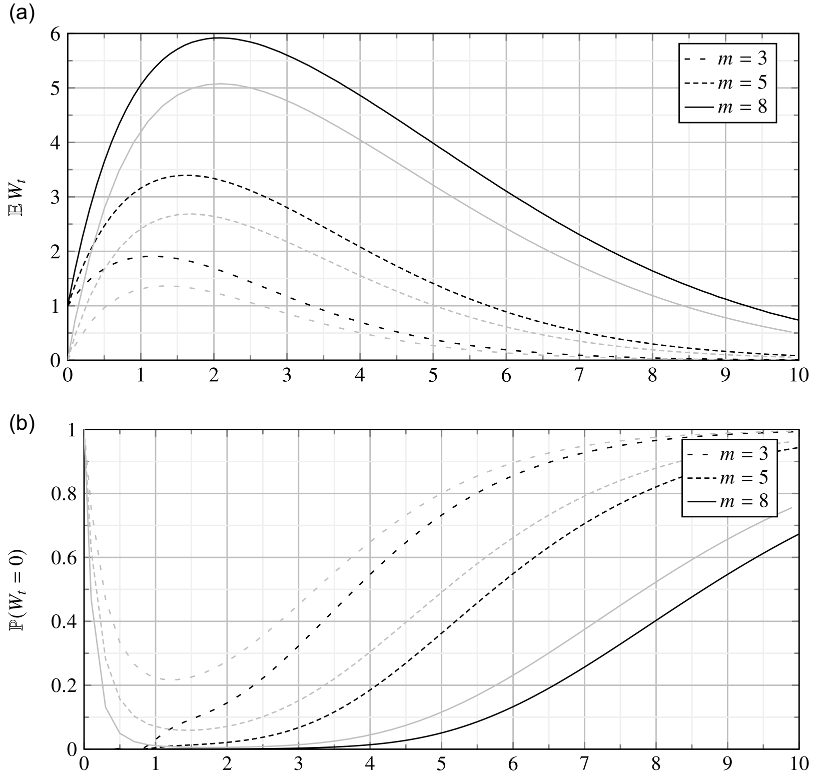

FIGURE 2. Mean workload (a) and empty-buffer probability (b) in the model with initial workload x, as functions of time, for different values of m and for $x = 1$. The lines in light gray denote the base model, in which case $x = 0$.

Combining the above findings, we have established the following result, which is numerically illustrated in Figure 2 (where the same instance is considered as in Figure 1).

Theorem 3.1. (Model with arbitrary initial workload.) *For any $\alpha \geq 0$ and $\beta > 0$, and $n \in \{0, \ldots, m\}$, the transform $\bar{\mathscr{G}}_n(\alpha, \beta \mid x, m)$ is given by*

$$\bar{\mathscr{G}}_n(\alpha, \beta \mid x, m) = \sum_{k=n}^{m} p_{m,k}(x, \beta)\, \bar{\mathscr{G}}_n(\alpha, \beta \mid 0, k) \\ + \mathrm{e}^{-\alpha x}\, \check{\mathscr{G}}_n(\alpha, \beta \mid m) - \sum_{k=n}^{m} p_{m,k}(x, \beta)\, \check{\mathscr{G}}_n(\alpha, \beta \mid k),$$

where (i) $\bar{\mathscr{G}}_n(\alpha, \beta \mid 0, k) = \beta\, \mathscr{G}_n(\alpha, \beta \mid k)$ *follows from Theorem* 2.1, (ii) $\check{\mathscr{G}}_n(\alpha, \beta \mid k)$ *can be found recursively from the equations* (3.3), *and* (iii) $p_{m,k}(x, \beta)$ *is given by* (3.4).

In the remainder of this section we present two immediate consequences of Theorem 3.1, both pertaining to the system starting empty at time 0: the first one describes the distribution of the first busy period, and the second the workload at an Erlang-distributed time epoch.

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

3.2. Busy period

In this subsection, we analyze the workload process's first busy period σ, which is distributed as $\sigma(B)$, with $m-1$ clients still to arrive from that point on.

The results of the previous subsection entail that

$$\begin{aligned}\mathbb{E}_m \mathrm{e}^{-\beta\sigma(x)} &= \mathbb{P}_m(\sigma(x) \leq T_\beta)\\ &= \sum_{k=0}^{m} \mathbb{P}_m(N(\sigma(x)) = k,\, \sigma(x) \leq T_\beta)\\ &= \sum_{k=0}^{m} \big(\exp(-\, VD(\beta)\, V^{-1}\, x)\big)_{m,k}\\ &= \sum_{k=0}^{m} (V \exp(-\, D(\beta)\, x) V^{-1})_{m,k}\\ &= \sum_{k=0}^{m} \gamma_{k,m} \mathrm{e}^{-d_k(\beta)\, x},\end{aligned}$$

for suitably chosen coefficients $\gamma_{k,m}$, with $k \in \{0, \dots, m\}$. As a consequence, recognizing the Laplace–Stieltjes transform of B, we find that

$$\mathbb{E}_m \mathrm{e}^{-\beta\sigma} = \int_0^\infty \sum_{k=0}^{m-1} \gamma_{k,m-1} \mathrm{e}^{-d_k(\beta)\, x}\, \mathbb{P}(B \in \mathrm{d}x) = \sum_{k=0}^{m-1} \gamma_{k,m-1} \mathscr{B}(d_k(\beta)).$$

3.3. Erlang horizon

In this subsection we point out how to compute the transform of $W(E_\beta(2))$ conditional on $W(0) = 0$, where $E_\beta(2)$ is an Erlang random variable with two phases and scale parameter β (or, put differently, $E_\beta(2)$ is distributed as the sum of two independent exponentially distributed random variables with mean β^{-1}).

Note that the quantity of our interest can be rewritten as

$$\mathbb{E}_{0,m}\big(\mathrm{e}^{-\alpha W(E_\beta(2))} \mathbf{1}_{\{N(E_\beta(2))=n\}}\big) = \int_0^\infty \sum_{k=n}^{m} \mathbb{P}_{0,m}(W(T_\beta) \in \mathrm{d}x,\, N(T_\beta) = k)\, \mathscr{G}_n(\alpha, \beta \mid x, k),$$

for $n \in \{0, \dots, m\}$. For suitably chosen coefficients $\bar{\gamma}_{\ell,k,n}$ (with $k \in \{n, \dots, m\}$ and $\ell \in \{n, \dots, k\}$) and $\check{\gamma}_{n,k}$ (with $k \in \{n, \dots, m\}$), by virtue of Theorem 3.1,

$$\mathscr{G}_n(\alpha, \beta \mid x, k) = \sum_{\ell=n}^{k} \bar{\gamma}_{\ell,k,n}\, \mathrm{e}^{-d_\ell(\beta)\, x} + \check{\gamma}_{n,k}\, \mathrm{e}^{-\alpha x}.$$

Upon combining the above observations,

$$\begin{aligned}&\mathbb{E}_{0,m}\big(\mathrm{e}^{-\alpha W(E_\beta(2))} \mathbf{1}_{\{N(E_\beta(2))=n\}}\big)\\ &\quad = \sum_{k=n}^{m} \sum_{\ell=n}^{k} \bar{\gamma}_{\ell,k,n}\, \bar{\mathscr{G}}_k(d_\ell(\beta), \beta \mid 0, m) + \sum_{k=n}^{m} \check{\gamma}_{n,k}\, \bar{\mathscr{G}}_k(\alpha, \beta \mid 0, m).\end{aligned}$$

This procedure extends in the obvious way to an Erlang horizon with more than two phases. Along the same lines, the joint distribution at time T_{β_1} and $T_{\beta_1}+T_{\beta_2}$, with T_{β_1} and T_{β_2} independent exponentially distributed random variables with means β_1^{-1} and β_2^{-1}, respectively, can be established.

4. External Poisson arrival stream

In this section we consider the case of an external Poisson stream of customers with i.i.d. service times. As before, our objective is to characterize the distribution of the transient workload in terms of Laplace–Stieltjes transforms.

The arrival rate of these external arrivals is $\bar{\lambda} \geq 0$, and the i.i.d. service times are distributed as the generic non-negative random variable $\bar{B}$ with cumulative distribution function $\bar{B}(\cdot)$ and Laplace–Stieltjes transform $\bar{\mathscr{B}}(\alpha) := \mathbb{E}\,\mathrm{e}^{-\alpha \bar{B}}$. Picking $\lambda = 0$ or $m = 0$, we have a conventional M/G/1 queue, whereas for $\bar{\lambda} = 0$ we recover the model of Section 2.

4.1. Recursion for the double transform

The counterpart of the partial differential equation (2.2) for this model with external Poisson arrivals is, for $x > 0$, $t > 0$ and $n \in \{0, \ldots, m-1\}$,

$$\begin{aligned}\frac{\partial}{\partial t}F_t(x,n) - f_t(x,n) = &-(\lambda n + \bar{\lambda})\,F_t(x,n) + \lambda(n+1)\int_{(0,x]} f_t(y,n+1)\,B(x-y)\,\mathrm{d}y \\ &+ \lambda(n+1)\,P_t(n+1)\,B(x) \\ &+ \bar{\lambda}\int_{(0,x]} f_t(y,n)\,\bar{B}(x-y)\,\mathrm{d}y + \bar{\lambda}\,P_t(n)\,\bar{B}(x),\end{aligned} \tag{4.1}$$

while for $n = m$ we have

$$\frac{\partial}{\partial t}F_t(x,m) - f_t(x,m) = -(\lambda m + \bar{\lambda})\,F_t(x,m) + \bar{\lambda}\int_{(0,x]} f_t(y,m)\,\bar{B}(x-y)\,\mathrm{d}y + \bar{\lambda}\,P_t(m)\,\bar{B}(x).$$

Multiplying (4.1) by $\mathrm{e}^{-\alpha x}\mathrm{e}^{-\beta t}$ and integrating over positive x and t, we obtain the algebraic equation

$$\begin{aligned}(\beta - \alpha)\mathscr{G}_n(\alpha,\beta) + \alpha\,\mathscr{P}_n(\beta) = &-(\lambda n + \bar{\lambda})\,\mathscr{G}_n(\alpha,\beta) + \lambda(n+1)\,\mathscr{G}_{n+1}(\alpha,\beta)\,\mathscr{B}(\alpha) \\ &+ \bar{\lambda}\,\mathscr{G}_n(\alpha,\beta)\,\bar{\mathscr{B}}(\alpha).\end{aligned} \tag{4.2}$$

Along the same lines,

$$(\beta - \alpha)\mathscr{G}_m(\alpha,\beta) + \alpha\,\mathscr{P}_m(\beta) - 1 = -(\lambda m + \bar{\lambda})\,\mathscr{G}_m(\alpha,\beta) + \bar{\lambda}\,\mathscr{G}_m(\alpha,\beta)\,\bar{\mathscr{B}}(\alpha). \tag{4.3}$$

4.2. Solving the double transform

As in Section 2, we start by finding an expression for $\mathscr{G}_m(\alpha,\beta)$. To this end, we isolate $\mathscr{G}_m(\alpha,\beta)$ in (4.3), so as to obtain

$$\mathscr{G}_m(\alpha,\beta) = \frac{\alpha\,\mathscr{P}_m(\beta) - 1}{\alpha - \beta - \lambda m - \bar{\lambda}(1 - \bar{\mathscr{B}}(\alpha))}. \tag{4.4}$$

The next step is to identify the unknown $\mathscr{P}_m(\beta)$. Define $\Phi(\alpha) := \alpha - \bar{\lambda}(1 - \bar{\mathscr{B}}(\alpha))$, in which we recognize the Laplace exponent of a compound Poisson process with drift, and $\Psi(\cdot)$ its right-inverse. Observing that the denominator of (4.4) vanishes when inserting $\alpha=\Psi(\beta+\lambda m)$, we conclude that $\mathscr{P}_m(\beta)=1/\Psi(\beta+\lambda m)$. This means that we have identified $\mathscr{G}_m(\alpha,\beta)$ as well:

$$\mathscr{G}_m(\alpha,\beta)=\frac{\Psi(\beta+\lambda m)-\alpha}{\beta+\lambda m-\Phi(\alpha)}\frac{1}{\Psi(\beta+\lambda m)}. \tag{4.5}$$

This is a familiar expression (see e.g. [6, Theorem 4.1]); compare the transform of the workload in an M/G/1 queue with arrival rate $\bar{\lambda}$ and service times distributed as $\bar{B}$, at an exponentially distributed time with mean $(\beta+\lambda m)^{-1}$.

We proceed by pointing out how $\mathscr{G}_0(\alpha,\beta),\ldots,\mathscr{G}_{m-1}(\alpha,\beta)$ can be found. We adopt the approach presented in Remark 2.3. For $n\in\{0,\ldots,m-1\}$, equation (4.2) leads to

$$\mathscr{G}_n(\alpha,\beta)=\frac{\alpha\mathscr{P}_n(\beta)-\lambda(n+1)\,\mathscr{G}_{n+1}(\alpha,\beta)\,\mathscr{B}(\alpha)}{\Phi(\alpha)-\beta-\lambda n}. \tag{4.6}$$

Observe that $\alpha_n := \Psi(\beta+\lambda n)$ is a root of the denominator, and hence a root of the numerator as well, leading to the equation

$$\mathscr{P}_n(\beta)=\frac{\lambda(n+1)\,\mathscr{G}_{n+1}(\Psi(\beta+\lambda n),\beta)\,\mathscr{B}(\Psi(\beta+\lambda n))}{\Psi(\beta+\lambda n)}; \tag{4.7}$$

cf. equation (2.9). Now the key idea, as in Remark 2.3, is to apply equations (4.7) and (4.6) alternately: inserting $n=m-1$ in (4.7) yields $\mathscr{P}_{m-1}(\beta)$, then inserting $n=m-1$ in (4.6) yields $\mathscr{G}_{m-1}(\alpha,\beta)$, and so on. We have established the following result.

Theorem 4.1. (Model with external Poisson arrival stream.) *For any $\alpha\geq 0$ and $\beta>0$, and $n\in\{0,\ldots,m-1\}$, the transform $\mathscr{G}_n(\alpha,\beta\mid m)$ is given by* (4.6), *where the transforms $\mathscr{P}_0(\beta),\ldots,\mathscr{P}_{m-1}(\beta)$ follow from recursion* (4.7), *with $\mathscr{P}_m(\beta)=1/\Psi(\beta+\lambda m)$, and $\mathscr{G}_m(\alpha,\beta\mid m)$ is given by* (4.5).

This result is numerically illustrated in Figure 3. As in the previous numerical experiments, we worked with exponentially distributed service times and $\lambda=\mu=1$. The service times of the external Poisson stream are exponentially distributed as well, with parameter $\bar{\mu}=5$.

5. Phase-type distributed arrival times

In Section 2 we saw that for exponentially distributed arrivals our model provides a closed-form solution, so it is a natural question whether the approach can be generalized to a more general class of arrival-time distributions. This class of distributions is particularly relevant, as any non-negative random variable can be approximated arbitrarily closely by a phase-type random variable [2, Section III.4]; the 'denseness' proof of [2, Theorem III.4.2] actually reveals that we can even restrict ourselves to a subclass of the phase-type distributions, namely the class of mixtures of Erlang distributions with different shape parameters but the same scale parameter. The proof of [2, Theorem III.4.2] also shows an intrinsic drawback of working with this specific class of phase-type distributions: one may need distributions of large dimension to get an accurate fit. This has motivated working with a low-dimensional two-moment fit, such as the one presented in [18]. In this fit, for distributions with a coefficient of variation less than 1, a mixture of two Erlang random variables (with the same scale parameter) is used, and for distributions with a coefficient of variation larger than 1, a hyperexponential random variable; for details see e.g. [16, Section 3.1].

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

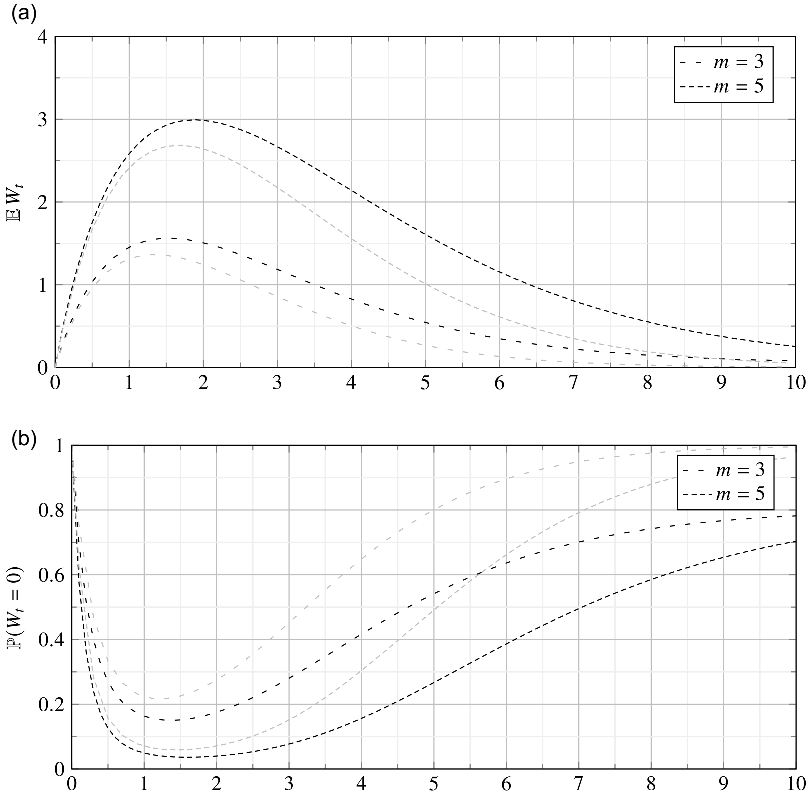

FIGURE 3. Mean workload (a) and empty-buffer probability (b), in the model with an external Poisson arrival stream, as functions of time, for different values of m. Here, the external Poisson arrival stream has rate parameter $\bar{\lambda} = 1$ and the service times of the customers are exponentially distributed with rate $\bar{\mu} = 5$. The lines in light gray denote the base model, in which case $\bar{\lambda} = 0$.

This section discusses the distribution of the transient workload in the case when the arrival times A are of phase type. An explicit expression in terms of transforms is still possible, albeit at the price of working with a large state space.

5.1. Set-up

We define the phase-type distribution of the generic arrival time A via the initial probability distribution $\boldsymbol{\gamma} \in \mathbb{R}^{d+1}$ (with $\gamma_i \geq 0$ for all phases i and $\boldsymbol{\gamma}^\top \mathbf{1} = 1$) and the transition matrix $Q = (q_{ij})_{i,j}^{d+1}$ (with non-negative off-diagonal elements and $Q\mathbf{1} = \mathbf{0}$); define $q_i := -q_{ii}$ and $\bar{q}_i := q_{i,d+1}$. The states 1 up to $d + 1$ are commonly referred to as the *phases* underlying the phase-type distribution. We assume that phase $d + 1$ is absorbing; from phase $i \in \{1, \ldots, d\}$ the process jumps to this phase with rate $\bar{q}_i$, after which the arrival takes place. As before, the m arrivals are independent of each other (and of the service times).

Let $\boldsymbol{N}(t)$ denote the state at time $t \geq 0$, in the sense that the ith component of $\boldsymbol{N}(t)$ denotes the number of the m clients that are in phase i at time t. It is clear that $\boldsymbol{N}_0$ has a multinomial

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

distribution with parameters m and $\gamma_1, \ldots, \gamma_{d+1}$, i.e. for any vector $\boldsymbol{n}_0 \in \mathbb{N}_0^{d+1}$ such that $\boldsymbol{n}_0^\top \mathbf{1} = m$,

$$\mathbb{P}(\boldsymbol{N}_0 = \boldsymbol{n}_0) = \binom{m}{n_{0,1}, \ldots, n_{0,d+1}} \prod_{i=1}^{d+1} \gamma_i^{n_{0,i}}.$$

Henceforth we therefore condition, without loss of generality, on the event $\{\boldsymbol{N}_0 = \boldsymbol{n}_0\}$; the transform of interest can be evaluated by deconditioning.

The key object of interest is the cumulative distribution function

$$F_t(x, \boldsymbol{n}) := \mathbb{P}(W(t) \le x, \boldsymbol{N}(t) = \boldsymbol{n}),$$

with $\boldsymbol{n} = (n_1, \ldots, n_{d+1}) \in \mathbb{N}_0^{d+1}$ such that $\boldsymbol{n}^\top \mathbf{1} = m$. In this section we work with the usual notation: the corresponding density is denoted by $f_t(x, \boldsymbol{n})$ for $x > 0$, while $P_t(\boldsymbol{n})$ is used to denote $\mathbb{P}(W(t) = 0, \boldsymbol{N}(t) = \boldsymbol{n})$. Mimicking the steps followed in Section 2, for any $x > 0$, up to $\mathrm{o}(\Delta t)$ terms,

$$\begin{aligned} F_{t+\Delta t}(x, \boldsymbol{n}) = &\left(1 - \sum_{i=1}^{d} n_i\, q_i \Delta t\right) F_t(x, \boldsymbol{n}) \\ &+ \sum_{i=1}^{d} \sum_{j\neq i}^{d} (n_i + 1) q_{ij} \Delta t\, \mathbf{1}_{\{n_j>0\}}\, F_t(x, \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_j) \\ &+ \sum_{i=1}^{d} (n_i + 1)\bar{q}_i \Delta t\, \mathbf{1}_{\{n_{d+1}>0\}} \Bigg(P_t(\boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_{d+1}) B(x) \\ &\qquad + \int_{(0,x]} f_t(y, \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_{d+1}) B(x - y)\, \mathrm{d}y \Bigg). \end{aligned} \tag{5.1}$$

This equation can be understood as follows. As in the exponential case dealt with in Section 2, the first term on the right-hand side corresponds to the scenario that no transitions between phases take place between t and $t + \Delta t$. The second term represents the transitions between two phases but not to the final phase. Finally, the third term covers a transition from one phase to the final phase, in which case the customer has arrived.

The next step is to convert equation (5.1) into a differential equation. After subtracting $F_t(x, \boldsymbol{n})$ from both sides of (5.1), dividing the full resulting equation by Δt, and letting $\Delta t \downarrow 0$, we arrive at the following partial differential equation:

$$\begin{aligned} \frac{\partial}{\partial t} F_t(x, \boldsymbol{n}) = &f_t(x, \boldsymbol{n}) - \sum_{i=1}^{d} n_i q_i F_t(x, \boldsymbol{n}) + \sum_{i=1}^{d} \sum_{j\neq i}^{d} (n_i + 1) q_{ij}\, \mathbf{1}_{\{n_j>0\}}\, F_t(x, \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_j) \\ &+ \sum_{i=1}^{d} (n_i + 1)\bar{q}_i\, \mathbf{1}_{\{n_{d+1}>0\}} \Bigg(P_t(\boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_{d+1}) B(x) \\ &\qquad + \int_{(0,x]} f_t(y, \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_{d+1}) B(x - y)\, \mathrm{d}y \Bigg). \end{aligned} \tag{5.2}$$

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

The partial differential equation (5.2) can be analyzed by transforming it twice, i.e. with respect to x and t: we first multiply (5.2) by $\mathrm{e}^{-\alpha x}$ and integrate x over $(0, \infty)$, and then we multiply the resulting equation by $\mathrm{e}^{-\beta t}$ and integrate t over $(0, \infty)$, where α and β are non-negative real numbers. To keep the resulting expressions as compact as possible, we will extensively work with the following objects:

$$\mathscr{F}_t(\alpha, \boldsymbol{n}) := \int_{(0,\infty)} \mathrm{e}^{-\alpha x} f_t(x, \boldsymbol{n})\, \mathrm{d}x, \qquad \mathscr{P}_{\boldsymbol{n}}(\beta) := \int_{(0,\infty)} \mathrm{e}^{-\beta t}\, P_t(\boldsymbol{n})\, \mathrm{d}t,$$

and

$$\bar{\mathscr{F}}_t(\alpha, \boldsymbol{n}) := P_t(\boldsymbol{n}) + \mathscr{F}_t(\alpha, \boldsymbol{n}), \qquad \mathscr{G}_{\boldsymbol{n}}(\alpha, \beta) := \int_{(0,\infty)} \mathrm{e}^{-\beta t}\, \bar{\mathscr{F}}_t(\alpha, \boldsymbol{n})\, \mathrm{d}t,$$

for $\alpha \geq 0$, $\beta > 0$, and, as before, $\mathscr{B}(a) = \mathbb{E}\, \mathrm{e}^{-\alpha B}$. After some tedious but elementary calculations we find that taking the double transform of (5.2) leads to the following result.

Lemma 5.1. *For any $\alpha \geq 0$ and $\beta > 0$, and $\boldsymbol{n} \in \mathbb{N}^{d+1}$ such that $\boldsymbol{n}^\top \mathbf{1} = m$,*

$$\begin{aligned}
&\left(\beta - \alpha + \sum_{i=1}^{d} n_i q_i\right) \mathscr{G}_{\boldsymbol{n}}(\alpha, \beta) \\
&\quad = \sum_{i=1}^{d} \sum_{j \neq i}^{d} (n_i + 1) q_{ij} \mathbf{1}_{\{n_j > 0\}} \mathscr{G}_{\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j}(\alpha, \beta) \\
&\qquad + \sum_{i=1}^{d} (n_i + 1) \bar{q}_i\, \mathbf{1}_{\{n_{d+1} > 0\}} \mathscr{B}(\alpha)\, \mathscr{G}_{\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_{d+1}}(\alpha, \beta) - \alpha \mathscr{P}_{\boldsymbol{n}}(\beta) + \mathbf{1}_{\{\boldsymbol{n}=\boldsymbol{n}_0\}}.
\end{aligned} \tag{5.3}$$

Proof. See Appendix B. □

We want to solve $\mathscr{G}_{\boldsymbol{n}}(\alpha, \beta)$ and $\mathscr{P}_{\boldsymbol{n}}(\beta)$ in (5.3) for all $\boldsymbol{n} \in \mathbb{N}_0^{d+1}$ such that $\boldsymbol{n}^\top \mathbf{1} = m$. Observe that for given $\mathscr{P}_{\boldsymbol{n}}(\beta)$, the functions $\mathscr{G}_{\boldsymbol{n}}(\alpha, \beta)$ follow by solving a system of linear equations.

5.2. Solution to the system of equations

Our next objective is to point out how the unknown functions $\mathscr{P}_{\boldsymbol{n}}(\beta)$ can be identified. To this end, first observe that we can determine $\mathscr{G}_{\boldsymbol{n}_0}(\alpha, \beta)$ analytically. We can also identify $\mathscr{P}_{\boldsymbol{n}}(\beta)$ for all $\boldsymbol{n}$ with $n_{d+1} = 0$, as $\mathscr{G}_{\boldsymbol{n}}(\alpha, \beta) = \mathscr{P}_{\boldsymbol{n}}(\beta)$ for these states, by solving the simplified system of equations in (5.3).

To find the remaining $\mathscr{G}_{\boldsymbol{n}}(\alpha, \beta)$ and $\mathscr{P}_{\boldsymbol{n}}(\beta)$ we will use an approach similar to that in Section 2.4. The state space of $\boldsymbol{N}(t)$ is the set of all configurations of m clients over $d + 1$ phases, so in total there are $\bar{m} := (m + d)!/(m!\, d!)$ states. Let $Q^{(\mathrm{Ph})}$ be the transition rate matrix of a continuous-time Markov process with rates

$$q^{(\mathrm{Ph})}_{\boldsymbol{n},\boldsymbol{n}'} = \begin{cases} n_i q_{ij} & \text{if } \boldsymbol{n}' = \boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_j, \\ 0 & \text{else.} \end{cases}$$

In addition, define the matrix $\mathscr{B}^{(\mathrm{Ph})}(\alpha)$ by

$$\left(\mathscr{B}^{(\mathrm{Ph})}(\alpha)\right)_{\boldsymbol{n},\boldsymbol{n}'} = 1 + (\mathscr{B}(\alpha) - 1)\, 1_{\{\boldsymbol{n}' = \boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_{d+1}\}}.$$

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

It takes some bookkeeping to verify that, for an appropriately chosen $\bar{m}$-dimensional vector $\boldsymbol{V}(\alpha, \beta)$ and $\bar{m} \times \bar{m}$ matrix $M(\alpha, \beta)$, the system of Lemma 5.1 can be rewritten in the form $M(\alpha, \beta)\,\boldsymbol{G}(\alpha, \beta) = \boldsymbol{V}(\alpha, \beta)$. Here $\boldsymbol{V}_{\boldsymbol{n}}(\alpha, \beta)$, i.e. the $\boldsymbol{n}$th entry of $\boldsymbol{V}(\alpha, \beta)$, is given by $\alpha\,\mathscr{P}_{\boldsymbol{n}}(\beta) - \boldsymbol{1}_{\{\boldsymbol{n}=\boldsymbol{n}_0\}}$, while $M(\alpha, \beta)$ denotes the transpose of the matrix exponent of a MAP, namely

$$(\alpha - \beta)I_{\bar{m}} + Q^{(\mathrm{Ph})} \circ \mathscr{B}^{(\mathrm{Ph})}(\alpha),$$

with $I_{\bar{m}}$ denoting an identity matrix of dimension $\bar{m}$, and $A \circ B$ denoting the Hadamard product of the matrices A and B. From this point on, the reasoning of Section 2.4 applies, with $\det M(\alpha, \beta)$ having $\bar{m}$ roots in the right-half of the complex α-plane (for any given $\beta > 0$), again by using [13] and [15]. This allows us to determine the unknown functions $\mathscr{P}_{\boldsymbol{n}}(\beta)$ by solving a system of linear equations; to see that these equations are linear, realize that $\det M_{\boldsymbol{n}}(\alpha, \beta)$, to be evaluated when applying Cramer's rule, depends linearly on the functions $\mathscr{P}_{\boldsymbol{n}}(\beta)$ (as appearing in the vector $\boldsymbol{V}(\alpha, \beta)$). As discussed in Section 2.4, if $\det M(\alpha, \beta) = \det M_{\boldsymbol{n}}(\alpha, \beta) = 0$, then also $\det M_{\boldsymbol{n}'}(\alpha, \beta) = 0$ for $\boldsymbol{n}' \neq \boldsymbol{n}$. Combining the above elements, we have found the following result.

Theorem 5.1. (Model with phase-type arrival times.) *For any $\alpha \geq 0$ and $\beta > 0$, and $\boldsymbol{n} \in \mathbb{N}^{d+1}$ such that $\boldsymbol{n}^\top \boldsymbol{1} = m$, the $\bar{m}$ transforms $\mathscr{G}_{\boldsymbol{n}}(\alpha, \beta)$ follow from the $\bar{m}$ linear equations (5.3). With $\alpha_1, \ldots, \alpha_{\bar{m}}$ the $\bar{m}$ solutions of $\det M(\alpha, \beta) = 0$ in the right-half of the complex α-plane, assumed to be distinct, the transforms $\mathscr{P}_{\boldsymbol{n}}(\beta)$ follow from the $\bar{m}$ linear equations $\det M_{\boldsymbol{n}}(\alpha_i, \beta) = 0$, with $i \in \{1, \ldots, \bar{m}\}$.*

5.3. A class with a straightforward solution

Above we pointed out how, in principle, the objects $\mathscr{G}_{\boldsymbol{n}}(\alpha, \beta)$ and $\mathscr{P}_{\boldsymbol{n}}(\beta)$, as appearing the system (5.3), can be found. However, it requires evaluation of determinants of large square matrices (of dimension $\bar{m} = (m + d)!/(m!\, d!)$). Fortunately, for important classes of phase-type distributions, we do not need to derive such determinants directly; instead, the system can be solved recursively. If the transition rate matrix underlying the phase-type distribution has no communicating states, then the phases can be rearranged such that $Q^{(\mathrm{Ph})}$ becomes upper triangular, and hence the corresponding $M(\alpha, \beta)$ as well. It means that the eigenvalues of $M(\alpha, \beta)$ are on its diagonal, and their roots are therefore of the form $\alpha_{\boldsymbol{n}} := \beta + \boldsymbol{q}^\top \boldsymbol{n}$. Following the same procedure as in Remark 2.3, we can alternately compute subsequent $\mathscr{G}_{\boldsymbol{n}}(\alpha, \beta)$ and $\mathscr{P}_{\boldsymbol{n}}(\beta)$. More concretely, first observe that

$$\mathscr{G}_{\boldsymbol{n}_0}(\alpha, \beta) = \mathscr{P}_{\boldsymbol{n}_0}(\beta) = \frac{1}{\alpha_{\boldsymbol{n}_0}} = \frac{1}{\beta + \boldsymbol{q}^\top \boldsymbol{n}_0}.$$

Then we can find an order of the states so that each $\mathscr{G}_{\boldsymbol{n}}(\alpha, \beta)$ can be evaluated based on its previously computed counterparts by applying (5.3), and

$$\mathscr{P}_{\boldsymbol{n}}(\beta) = \frac{1}{\alpha_{\boldsymbol{n}}} \left(\sum_{i=1}^{d} \sum_{j \neq i}^{d} (n_i + 1) q_{ij} \boldsymbol{1}_{\{n_j > 0\}} \mathscr{G}_{\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j}(\alpha_{\boldsymbol{n}}, \beta) \right.$$
$$\left. + \sum_{i=1}^{d} (n_i + 1)\bar{q}_i\, \boldsymbol{1}_{\{n_{d+1} > 0\}} \mathscr{B}(\alpha_{\boldsymbol{n}})\, \mathscr{G}_{\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_{d+1}}(\alpha_{\boldsymbol{n}}, \beta) + \boldsymbol{1}_{\{\boldsymbol{n}=\boldsymbol{n}_0\}} \right).$$

In the remainder of this subsection we detail this procedure for two frequently used phase-type distributions.

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

5.3.1. *Erlang-distributed arrival times.* In this section we consider the case when A has an Erlang distribution, characterized by the shape parameter $k \in \mathbb{N}$ and the scale parameter $\lambda > 0$. An attractive feature of the Erlang random variable is that it is equivalent to the sum of k independent exponential variables with parameter λ, which allows us to implement it as a phase-type distribution. We can represent each arrival as consisting of k exponentially distributed phases. Note that $\boldsymbol{N}(t) \in \mathbb{N}_0^{k+1}$, with $\boldsymbol{N}(t)^\top \mathbf{1} = m$. All customers start in the first phase, i.e. $\boldsymbol{N}_0 = \boldsymbol{n}_0 = (m, 0, \dots, 0)$. A transition from the first to the second entry of $\boldsymbol{N}(t)$ (i.e. first entry going down by one, second going up by one) represents the first exponential random variable in the Erlang distribution, a transition from the second to the third entry represents the second exponential random variable, and so on. As transitions can only happen to the next phase, there is no communication between phases. When the final transition happens, i.e. from the kth entry of $\boldsymbol{N}(t)$ to the $(k+1)$th, the arrival of the customer into the system has taken place, and the workload increases.

By applying (5.3), we are to solve the system of equations given by

$$\mathscr{G}_{\boldsymbol{n}}(\alpha, \beta) = \frac{1}{\beta - \alpha + \lambda \sum_{i=1}^{k} n_i} \Bigg(\sum_{i=1}^{k-1} (n_i + 1)\lambda\, \mathbf{1}_{\{n_{i+1} > 0\}} \mathscr{G}_{\boldsymbol{n}+\boldsymbol{e}_i - \boldsymbol{e}_{i+1}}(\alpha, \beta) \\ + (n_k + 1)\lambda\, \mathbf{1}_{\{n_{k+1} > 0\}} \mathscr{B}(\alpha) \mathscr{G}_{\boldsymbol{n}+\boldsymbol{e}_k - \boldsymbol{e}_{k+1}}(\alpha, \beta) - \alpha \mathscr{P}_{\boldsymbol{n}}(\beta) + \mathbf{1}_{\{\boldsymbol{n}=\boldsymbol{n}_0\}} \Bigg). \tag{5.4}$$

Since $\mathscr{G}_{\boldsymbol{n}}(\alpha, \beta) = \mathscr{P}_{\boldsymbol{n}}(\beta)$ for all $\boldsymbol{n}$ such that $n_{k+1} = 0$, states in which no arrivals have taken place yet, all $\mathscr{G}_{\boldsymbol{n}}(\alpha, \beta)$ can be determined from the system in (5.4) for these $\boldsymbol{n}$. Considering the initialization of the recursion, i.e. $\boldsymbol{n} = \boldsymbol{n}_0$, we have

$$\mathscr{P}_{\boldsymbol{n}_0}(\beta) = \mathscr{G}_{\boldsymbol{n}_0}(\alpha, \beta) = \int_0^\infty \mathrm{e}^{-\beta t} P_t(\boldsymbol{n}_0)\, \mathrm{d}t = \int_0^\infty \mathrm{e}^{-\beta t} \mathrm{e}^{-\lambda m t}\, \mathrm{d}t = \frac{1}{\beta + \lambda m}.$$

To demonstrate the procedure, we proceed by determining the first few $\mathscr{G}_{\boldsymbol{n}}(\alpha, \beta)$ with $n_{k+1} = 0$. With $\boldsymbol{n}^{(1)} := (m-1, 1, 0, \dots, 0)$, equation (5.4) gives

$$\mathscr{G}_{\boldsymbol{n}^{(1)}}(\alpha, \beta) = \frac{m\lambda \mathscr{G}_{\boldsymbol{n}_0}(\alpha, \beta) - \alpha \mathscr{G}_{\boldsymbol{n}^{(1)}}(\alpha, \beta)}{\beta - \alpha + \lambda m},$$

so that

$$\mathscr{G}_{\boldsymbol{n}^{(1)}}(\alpha, \beta) = \frac{\lambda m}{\beta + \lambda m} \mathscr{G}_{\boldsymbol{n}_0}(\alpha, \beta) = \frac{\lambda m}{(\beta + \lambda m)^2}.$$

Similarly, for $\boldsymbol{n}^{(2)} := (m-1, 0, 1, 0, \dots, 0)$ and $\boldsymbol{n}^{(3)} := (m-2, 2, 0, \dots, 0)$ we find, respectively,

$$\mathscr{G}_{\boldsymbol{n}^{(2)}}(\alpha, \beta) = \frac{\lambda}{\beta + \lambda m} \mathscr{G}_{\boldsymbol{n}^{(1)}}(\alpha, \beta) = \frac{\lambda^2 m}{(\beta + \lambda m)^3}$$

and

$$\mathscr{G}_{\boldsymbol{n}^{(3)}}(\alpha, \beta) = \frac{\lambda(m-1)}{\beta + \lambda m} \mathscr{G}_{\boldsymbol{n}^{(1)}}(\alpha, \beta) = \frac{\lambda^2 m(m-1)}{(\beta + \lambda m)^3}.$$

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

5.3.2. *Hyperexponentially distributed arrival times.* We conclude this section by discussing the case when A is hyperexponentially distributed. We assume that A is defined via $k \in \mathbb{N}$ exponentially distributed random variables; the ith, with its own parameter $\lambda_i > 0$, is picked with probability p_i (where $\boldsymbol{p}^\top \mathbf{1} = 1$), for $i \in \{1, \dots, k\}$. To represent this as a phase-type distributed arrival, we set the dimension of $\boldsymbol{N}(t)$ equal to $k + 1$ (i.e. one phase for each type of exponentially distributed random variables, plus the absorbing state). We generate a starting state $\boldsymbol{n}_0$ according to a multinomial distribution with parameters m and p_i (clearly $p_{k+1} = 0$). Throughout the following analysis we condition on the event $\{\boldsymbol{N}_0 = \boldsymbol{n}_0\}$, entailing that we draw the type of each of the customers beforehand. Note that all customers make a transition from their current phase directly to phase $k + 1$, after which an arrival takes place.

By applying (5.3), we obtain the following system of equations:

$$\mathscr{G}_{\boldsymbol{n}}(\alpha, \beta) = \frac{\sum_{i=1}^{k} (n_i + 1)\mathbf{1}_{\{n_i < n_{0,i}\}}\lambda_i \mathscr{B}(\alpha)\mathscr{G}_{\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_{k+1}}(\alpha, \beta) - \alpha \mathscr{P}_{\boldsymbol{n}}(\beta) + \mathbf{1}_{\{\boldsymbol{n}=\boldsymbol{n}_0\}}}{\beta - \alpha + \boldsymbol{\lambda}^\top \boldsymbol{n}}.$$

Again, we initialize for $\mathscr{G}_{\boldsymbol{n}_0}(\alpha, \beta)$ by a direct computation:

$$\mathscr{P}_{\boldsymbol{n}_0}(\beta) = \mathscr{G}_{\boldsymbol{n}_0}(\alpha, \beta) = \int_0^\infty \mathrm{e}^{-\beta t} P_t(\boldsymbol{n}_0)\,\mathrm{d}t = \int_0^\infty \mathrm{e}^{-\beta t}\mathrm{e}^{-\boldsymbol{\lambda}^\top \boldsymbol{n}_0 t}\,\mathrm{d}t = \frac{1}{\beta + \boldsymbol{\lambda}^\top \boldsymbol{n}_0}.$$

6. Balking customers

In this section, each arriving customer decides, based on the workload seen upon arrival, whether or not they join the queue. In queueing theory, this mechanism is often referred to as *balking*: the higher the workload, the less the customer is inclined to join. In the model considered we work with an exponential balking distribution: if the current workload is smaller than an exponentially random variable with mean θ^{-1}, the customer enters the system.

We start by setting up the counterpart of the partial differential equation (2.1), relying on the methodology that we have been using before. First observe that, as $\Delta t \downarrow 0$, for any $x > 0$, $t > 0$, and $n \in \{0, \dots, m-1\}$,

$$\begin{aligned} &F_{t+\Delta t}(x, n) \\ &\quad = (1 - \lambda n\,\Delta t) \cdot F_t(x + \Delta t, n) + \lambda(n+1)\,\Delta t \int_{(0,x]} f_t(y, n+1)\,\mathrm{e}^{-\theta y}\,B(x-y)\,\mathrm{d}y \\ &\qquad + \lambda(n+1)\,\Delta t\,P_t(n+1)\,B(x) + \lambda(n+1)\,\Delta t \int_{(0,x]} f_t(y, n+1)\,(1 - \mathrm{e}^{-\theta y})\,\mathrm{d}y + \mathrm{o}(\Delta t). \end{aligned}$$

In the second term on the right-hand side, the factor $\mathrm{e}^{-\theta y}$ represents the probability that the customer joins if they are facing a workload y, and in the fourth term the factor $1 - \mathrm{e}^{-\theta y}$ represents the probability that the customer does not join if they are facing a workload y. In the usual manner, this leads to the partial differential equation

$$\begin{aligned} \frac{\partial}{\partial t}F_t(x, n) - f_t(x, n) = &- \lambda n\,F_t(x, n) + \lambda(n+1)\int_{(0,x]} f_t(y, n+1)\,\mathrm{e}^{-\theta y}\,B(x-y)\,\mathrm{d}y \\ &+ \lambda(n+1)\,P_t(n+1)\,B(x) + \lambda(n+1)\int_{(0,x]} f_t(y, n+1)\,(1 - \mathrm{e}^{-\theta y})\,\mathrm{d}y. \end{aligned}$$

We follow the same procedure as before: we transform subsequently to x and t. This means that we first multiply the previous display by $\mathrm{e}^{-\alpha x}$ and integrate over positive x. Using calculations

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

similar to those used in Section 2, and splitting $\mathrm{e}^{-\alpha x} = \mathrm{e}^{-\alpha y}\mathrm{e}^{-\alpha(x-y)}$, we obtain the following ordinary differential equation:

$$\begin{aligned}\frac{\partial}{\partial t}\frac{P_t(n)+\mathscr{F}_t(\alpha, n)}{\alpha} - \mathscr{F}_t(\alpha, n) = -\lambda n\,\frac{P_t(n)+\mathscr{F}_t(\alpha, n)}{\alpha}\\ + \lambda(n+1)\,\frac{(P_t(n+1)+\mathscr{F}_t(\alpha+\theta, n+1))\,\mathscr{B}(\alpha)}{\alpha}\\ + \lambda(n+1)\,\frac{\mathscr{F}_t(\alpha, n+1)-\mathscr{F}_t(\alpha+\theta, n+1)}{\alpha}.\end{aligned}$$

Then we transform with respect to time: we multiply by $\mathrm{e}^{-\beta t}$ and integrate over positive t. We thus find, for $n \in \{0, \ldots, m-1\}$,

$$\begin{aligned}(\beta-\alpha)\mathscr{G}_n(\alpha, \beta) + \alpha\,\mathscr{P}_n(\beta) = -\lambda n\,\mathscr{G}_n(\alpha, \beta) + \lambda(n+1)\,\mathscr{G}_{n+1}(\alpha+\theta, \beta)\,\mathscr{B}(\alpha)\\ + \lambda(n+1)\,(\mathscr{G}_{n+1}(\alpha, \beta) - \mathscr{G}_{n+1}(\alpha+\theta, \beta)).\end{aligned} \tag{6.1}$$

It directly follows from equation (6.1) that

$$\mathscr{G}_n(\alpha, \beta) = \frac{\lambda(n+1)\,(\mathscr{G}_{n+1}(\alpha, \beta) - (1-\mathscr{B}(\alpha))\mathscr{G}_{n+1}(\alpha+\theta, \beta)) - \alpha\,\mathscr{P}_n(\beta)}{\beta-\alpha+\lambda n}. \tag{6.2}$$

With $\alpha_n := \beta + \lambda n$ as before, the usual argumentation yields that

$$\mathscr{P}_n(\beta) = \lambda(n+1)\frac{\mathscr{G}_{n+1}(\alpha_n, \beta) - (1-\mathscr{B}(\alpha_n))\mathscr{G}_{n+1}(\alpha_n+\theta, \beta)}{\alpha_n}. \tag{6.3}$$

Because we know that

$$\mathscr{P}_m(\beta) = \mathscr{G}_m(\alpha, \beta) = \frac{1}{\beta+\lambda m}, \tag{6.4}$$

all transforms involved can be recursively identified following the recipe discussed in Remark 2.3, i.e. by applying equations (6.2) and (6.3) alternately. We have thus found the following result, numerically illustrated in Figure 4 (again with exponentially distributed service times, and $\lambda = \mu = 1$).

Theorem 6.1. (Model with balking customers.) *For any $\alpha \geq 0$ and $\beta > 0$, and $n \in \{0, \ldots, m-1\}$, the transform $\mathscr{G}_n(\alpha, \beta \mid m)$ is given by (6.2), where the transforms $\mathscr{P}_0(\beta), \ldots, \mathscr{P}_{m-1}(\beta)$ follow from (6.3), and $\mathscr{P}_m(\beta)$ and $\mathscr{G}_m(\alpha, \beta \mid m)$ are given by (6.4).*

Remark 6.1. Above we assumed an exponential balking distribution, but the procedure naturally extends to more general distributions. Hyperexponential balking is straightforward to deal with, whereas for Erlang balking the procedure is slightly more delicate. For instance, for an Erlang distribution with shape parameter 2 and scale parameter θ, recalling that $(1+\theta y)\mathrm{e}^{-\theta y}$ is the probability that this random variable is larger than y, one has, where we locally abbreviate $f(x) \equiv f_t(x, n+1)$ and $\mathscr{F}(\alpha) \equiv \int_{(0,\infty)} \mathrm{e}^{-\alpha x} f(x)\,\mathrm{d}x$,

$$\int_0^\infty \mathrm{e}^{-\alpha x}\int_{(0,x]} f(y)\,(1+\theta y)\,\mathrm{e}^{-\theta y}\,B(x-y)\,\mathrm{d}y\,\mathrm{d}x = (\mathscr{F}(\alpha+\theta) - \theta\,\mathscr{F}'(\alpha+\theta))\,\frac{\mathscr{B}(\alpha)}{\alpha},$$

which can then be transformed with respect to t in the usual manner. In general, if the scale parameter of the Erlang balking is $k \in \mathbb{N}$, kth-order derivatives of $\mathscr{F}$ will appear. Observe that the proposed recursive procedure can still be performed.

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

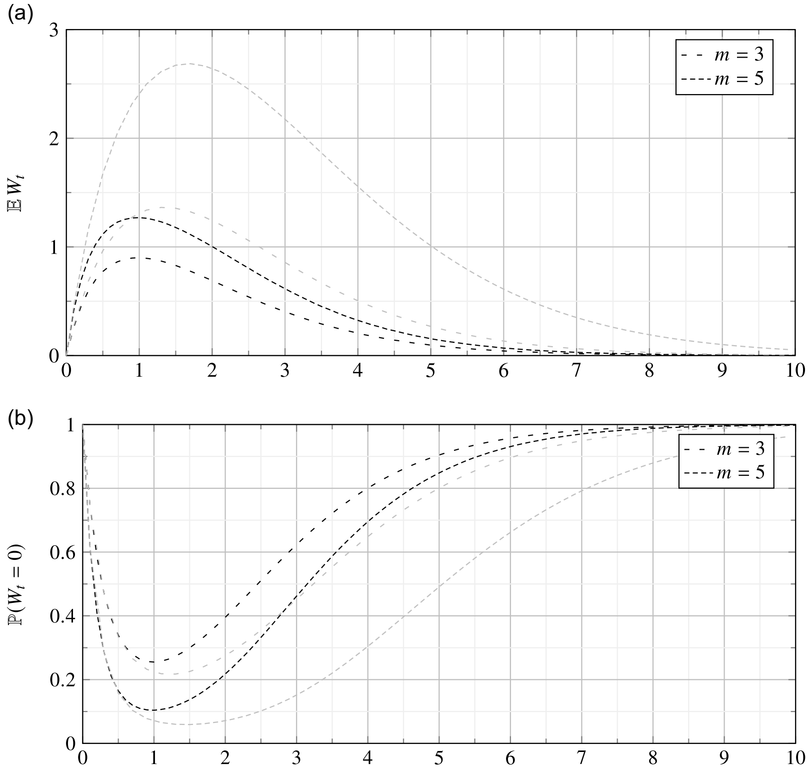

FIGURE 4. Mean workload (a) and empty-buffer probability (b), as functions of time in the model with balking customers, for different values of m and for $\theta = 11/10$. The lines in light gray denote the base model, in which case $\theta = 0$.

Remark 6.2. This section has considered a model with a finite pool of potentially joining customers (as was the case in the other sections of this paper). Interestingly, mimicking the analysis of the present section reveals that the corresponding model with an *infinite pool* (i.e. the model in which customers arrive according to a Poisson process) does not lead to a clean solution, the reason being that we arrive at an equation in which both a transform evaluated in α and the same transform evaluated in $\alpha + \theta$ appear. This indicates that this is an example of a model in which the finite-population version is easier to handle than its infinite-population counterpart (in which the relevant transformed could be recursively solved; see Theorem 6.1).

A similar phenomenon happens in the model with a finite customer population, balking, and retrial: each customer who has decided not to join then retries after an exponentially distributed time with parameter λ. Equation (6.1) thus becomes

$$(\beta - \alpha)\mathscr{G}_n(\alpha, \beta) + \alpha\, \mathscr{P}_n(\beta) = -\lambda n\, \mathscr{G}_n(\alpha, \beta) + \lambda(n+1)\, \mathscr{G}_{n+1}(\alpha + \theta, \beta)\, \mathscr{B}(\alpha) \\ + \lambda n\, (\mathscr{G}_n(\alpha, \beta) - \mathscr{G}_n(\alpha + \theta, \beta)),$$

with the same intrinsic difficulty.

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

7. Discussion and concluding remarks

This paper has considered the workload in a queue with a finite number of independently arriving customers. We thus deviate from the conventional queueing paradigm in which there is an infinite pool of customers, who request service with independent interarrival times. We subsequently consider a base model in which the arrival times are exponentially distributed, and various extensions. These extensions cover the settings (a) in which the initial workload level has an arbitrary non-negative value, (b) with an additional external Poisson arrival stream, (c) with phase-type arrival times, and (d) with balking customers.

In existing work, various asymptotic regimes have been considered; see e.g. the diffusion scaling of [4] and the large deviations of [10]. They give rise to two particularly interesting research questions.

- Can we identify the tail asymptotics of the transient workload $W(t)$, or the workload faced by the nth customer? Importantly, techniques that have been used for the case of independent interarrival times, in the context of GI/G/1 queues, cannot be applied here. Most notably, while one could in principle convert the tail probabilities into those of an associated random walk, this random walk has no i.i.d. ladder heights. If the target is to identify logarithmic asymptotics in the regime when the service times B are light-tailed, the results of [10] can potentially be applied.

- In our framework the server's processing speed is constant over time (i.e. equal to one). One could try to extend the analysis to a time-dependent service speed $c(t)$, so as to model the service system's staffing profile. A relevant question is as follows: Can we develop ways to determine a staffing profile such that, uniformly in time, a certain minimum performance criterion is maintained? While it seems highly challenging to perform an exact analysis, there could be openings in asymptotic regimes.

Appendix A. Proof of Lemma 2.1

In this section we show how we derived the determinant of the matrix M_n. By construction, we know that M_n is of the following form:

$$\begin{bmatrix}
a_0 & b_0 & 0 & & 0 & c_0 & 0 & & 0 & 0 & 0 \\
0 & a_1 & b_1 & \dots & 0 & c_1 & 0 & \dots & 0 & 0 & 0 \\
0 & 0 & a_2 & & 0 & c_2 & 0 & & 0 & 0 & 0 \\
 & \vdots & & \ddots & & & & & & \vdots & \\
0 & 0 & 0 & & a_{n-1} & c_{n-1} & 0 & & 0 & 0 & 0 \\
0 & 0 & 0 & & 0 & c_n & b_n & & 0 & 0 & 0 \\
0 & 0 & 0 & & 0 & c_{n+1} & a_{n+1} & & 0 & 0 & 0 \\
 & \vdots & & & & & & \ddots & & \vdots & \\
0 & 0 & 0 & & 0 & c_{m-3} & 0 & & a_{m-3} & b_{m-3} & 0 \\
0 & 0 & 0 & \dots & 0 & c_{m-2} & 0 & \dots & 0 & a_{m-2} & b_{m-2} \\
0 & 0 & 0 & & 0 & c_{m-1} & 0 & & 0 & 0 & a_{m-1}
\end{bmatrix}. \tag{A.1}$$

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

This matrix is very close to being an upper triangular matrix, for which the determinant is simply the product of the entries on the diagonal. We can transform (A.1) to an upper triangular matrix by elementary column operations, under which the determinant does not change. We achieve this by applying the following algorithm.

(1) From column n subtract c_{m-1}/a_{m-1} times column $m-1$. This results in the last entry of column n becoming 0, while the $(m-2)$th entry becomes

$$c^\star_{m-2} = c_{m-2} - \frac{c_{m-1}}{a_{m-1}}\, b_{m-2}.$$

(2) For $i = m-2, \ldots, n+1$, from column n subtract $c^\star_i/a_i$ times column i. The ith entry of column n becomes 0, while the $(i-1)$th entry becomes

$$c^\star_{i-1} = c_{i-1} - \frac{c^\star_i}{a_i}\, b_{i-1}.$$

This algorithm results in an upper triangular matrix with the (n, n)th entry being $c^\star_n$, which is given by the recursion

$$c^\star_n = c_n - \frac{b_n}{a_{n+1}} c^\star_{n+1}, \qquad c^\star_{m-1} = c_{m-1},$$

which is solved by

$$c^\star_n = c_{m-1} \prod_{i=n}^{m-2}\left(-\frac{b_i}{a_{i+1}}\right) + \sum_{j=n}^{m-2} c_j \prod_{i=n}^{j-1}\left(-\frac{b_i}{a_{i+1}}\right).$$

The entries of the matrix M_n are given by $a_i = -(\beta - \alpha + \lambda i)$, $b_i = \lambda(i+1)\mathscr{B}(\alpha)$, $c_{m-1} = \alpha\mathscr{P}_{m-1}(\beta) - \lambda m\,\mathscr{B}(\alpha)\mathscr{P}_m(\beta)$, and $c_i = \alpha\mathscr{P}_i(\beta)$ for $i \neq m-1$. Substituting these yields

$$\begin{aligned}
c^\star_n &= (\alpha\mathscr{P}_{m-1}(\beta) - \lambda m\mathscr{B}(\alpha)\mathscr{P}_m(\beta)) \prod_{i=n}^{m-2} \frac{\lambda(i+1)\mathscr{B}(\alpha)}{\beta - \alpha + \lambda(i+1)} + \sum_{j=n}^{m-2} \alpha\mathscr{P}_j(\beta) \prod_{i=n}^{j-1} \frac{\lambda(i+1)\mathscr{B}(\alpha)}{\beta - \alpha + \lambda(i+1)} \\
&= \alpha\mathscr{P}_n(\beta) - \lambda m\mathscr{B}(\alpha)\mathscr{P}_m(\beta) \frac{(\lambda\mathscr{B}(\alpha))^{m-(n+1)}(m-1)!/n!}{\xi_{m-1}/\xi_n} \\
&\quad + \sum_{j=n+1}^{m-1} \alpha\mathscr{P}_j(\beta) \frac{(\lambda\mathscr{B}(\alpha))^{j-n} j!/n!}{\xi_j/\xi_n} \\
&= \alpha\mathscr{P}_n(\beta) - \frac{(\lambda\mathscr{B}(\alpha))^{m-n} m!\, \xi_n}{n!\, \xi_{m-1}} \mathscr{P}_m(\beta) + \frac{(\lambda\mathscr{B}(\alpha))^{-n}\xi_n}{n!\, \xi_{m-1}} \sum_{j=n+1}^{m-1} \alpha\mathscr{P}_j(\beta) \frac{(\lambda\mathscr{B}(\alpha))^j j!\, \xi_{m-1}}{\xi_j} \\
&= \alpha\mathscr{P}_n(\beta) - \frac{1}{(\lambda\mathscr{B}(\alpha))^n n!\, (\xi_{m-1}/\xi_n)} \left((\lambda\mathscr{B}(\alpha))^m m!\, \mathscr{P}_m(\beta) - \sum_{j=n+1}^{m-1} (-\mathscr{H}_j(\alpha, \beta))\mathscr{P}_j(\beta) \right)
\end{aligned}$$

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

$$= \alpha \mathscr{P}_n(\beta) + \frac{\alpha}{\mathscr{H}_n(\alpha,\beta)}\left(\mathscr{H}_m(\alpha,\beta) + \sum_{j=n+1}^{m-1} \mathscr{H}_j(\alpha,\beta)\mathscr{P}_j(\beta)\right)$$

$$= \alpha\left(\mathscr{P}_n(\beta) + \frac{\mathscr{H}_m(\alpha,\beta) + \sum_{j=n+1}^{m-1} \mathscr{H}_j(\alpha,\beta)\mathscr{P}_j(\beta)}{\mathscr{H}_n(\alpha,\beta)}\right),$$

as desired. □

Appendix B. Proof of Lemma 5.1

We transform (5.2) by multiplying it by $\mathrm{e}^{-\alpha x}$ and integrating x over $(0, \infty)$, for $\alpha \geq 0$. Applying integration of parts,

$$\int_{(0,\infty)} \mathrm{e}^{-\alpha x} F_t(x, \boldsymbol{n})\,\mathrm{d}x = \left[-\frac{\mathrm{e}^{-\alpha x}}{\alpha} F_t(x,\boldsymbol{n})\right]_0^\infty + \frac{1}{\alpha}\int_{(0,\infty)} \mathrm{e}^{-\alpha x} f_t(x,\boldsymbol{n})\,\mathrm{d}x$$
$$= \frac{P_t(\boldsymbol{n}) + \mathscr{F}_t(\alpha,\boldsymbol{n})}{\alpha}$$
$$= \frac{\bar{\mathscr{F}}_t(\alpha,\boldsymbol{n})}{\alpha},$$

which also gives, as a consequence of Leibniz's integral rule,

$$\int_{(0,\infty)} \mathrm{e}^{-\alpha x} \frac{\partial}{\partial t} F_t(x,\boldsymbol{n})\,\mathrm{d}x = \frac{\partial}{\partial t}\int_{(0,\infty)} \mathrm{e}^{-\alpha x} F_t(x,\boldsymbol{n})\,\mathrm{d}x = \frac{\partial}{\partial t}\frac{\bar{\mathscr{F}}_t(\alpha,\boldsymbol{n})}{\alpha}.$$

Furthermore, by Fubini's theorem,

$$\int_{(0,\infty)} \mathrm{e}^{-\alpha x}\int_{(0,x]} f_t(y, \boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j)\,\mathrm{d}y\,\mathrm{d}x = \int_{(0,\infty)}\int_{(y,\infty)} \mathrm{e}^{-\alpha x} f_t(y,\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j)\,\mathrm{d}x\,\mathrm{d}y$$
$$= -\int_{(0,\infty)} \frac{\mathrm{e}^{-\alpha y}}{\alpha} f_t(y,\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j)\,\mathrm{d}y$$
$$= -\frac{\mathscr{F}_t(\alpha,\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j)}{\alpha},$$

and, again by Fubini and integration by parts,

$$\int_{(0,\infty)} \mathrm{e}^{-\alpha x}\int_{(0,x]} f_t(y,\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_{d+1})B(x-y)\,\mathrm{d}y\,\mathrm{d}x$$
$$= \int_{(0,\infty)} f_t(y,\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_{d+1})\int_{(y,\infty)} \mathrm{e}^{-\alpha x}B(x-y)\,\mathrm{d}x\,\mathrm{d}y$$
$$= \int_{(0,\infty)} f_t(y,\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_{d+1})\left(\left[-\frac{\mathrm{e}^{-\alpha x}}{\alpha}B(x-y)\right]_y^\infty + \int_{(y,\infty)} \frac{\mathrm{e}^{-\alpha x}}{\alpha}\frac{\partial}{\partial x}B(x-y)\,\mathrm{d}x\right)\mathrm{d}y$$
$$= \frac{1}{\alpha}\int_{(0,\infty)} f_t(y,\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_{d+1})\int_{(0,\infty)} \mathrm{e}^{-\alpha u}\mathrm{e}^{-\alpha y}\frac{\partial}{\partial u}B(u)\,\mathrm{d}u\,\mathrm{d}y$$

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

$$
\begin{aligned}
&= \frac{\mathscr{B}(\alpha)}{\alpha} \int_{(0,\infty)} \mathrm{e}^{-\alpha y} f_t(y, \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_{d+1}) \, \mathrm{d}y \\
&= \frac{\mathscr{B}(\alpha) \, \mathscr{F}_t(\alpha, \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_{d+1})}{\alpha};
\end{aligned}
$$

here we have used $B(0) = 0$, a simple substitution $u = x - y$, and the notation $\mathscr{B}(\alpha) = \mathbb{E}\, \mathrm{e}^{-\alpha B}$. For the final two terms, we obtain

$$
\int_{(0,\infty)} \mathrm{e}^{-\alpha x} P_t(\boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_j) \, \mathrm{d}x = \frac{P_t(\boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_j)}{\alpha},
$$

and

$$
\begin{aligned}
&\int_{(0,\infty)} \mathrm{e}^{-\alpha x} P_t(\boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_{d+1}) B(x) \, \mathrm{d}x \\
&\quad = P_t(\boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_{d+1}) \left(\left[-\frac{\mathrm{e}^{-\alpha x}}{\alpha} B(x) \right]_0^\infty + \int_{(0,\infty)} \frac{\mathrm{e}^{-\alpha x}}{\alpha} \frac{\partial}{\partial x} B(x) \, \mathrm{d}x \right) \\
&\quad = \frac{\mathscr{B}(\alpha) P_t(\boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_{d+1})}{\alpha}.
\end{aligned}
$$

Putting everything together and multiplying by α gives

$$
\begin{aligned}
&\frac{\partial}{\partial t} \bar{\mathscr{F}}_t(\alpha, \boldsymbol{n}) - \alpha(\bar{\mathscr{F}}_t(\alpha, \boldsymbol{n}) - P_t(\boldsymbol{n})) \\
&\quad = -\sum_{i=1}^{d} n_i q_i \bar{\mathscr{F}}_t(\alpha, \boldsymbol{n}) + \sum_{i=1}^{d} \sum_{j \neq i, d+1} (n_i + 1) q_{ij} \mathbf{1}_{\{n_j > 0\}} \bar{\mathscr{F}}_t(\alpha, \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_j) \\
&\qquad + \sum_{i=1}^{d} (n_i + 1) \bar{q}_i \, \mathbf{1}_{\{n_{d+1} > 0\}} \mathscr{B}(a) \bar{\mathscr{F}}_t(\alpha, \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_{d+1}).
\end{aligned}
\tag{B.1}
$$

So as to perform the second transform, we multiply (B.1) by $\mathrm{e}^{-\beta t}$, for $\beta > 0$, and integrate t over $(0, \infty)$. All terms are straightforward, except for the first term, for which again we need to use integration by parts. We obtain

$$
\begin{aligned}
\int_{(0,\infty)} \mathrm{e}^{-\beta t} \frac{\partial}{\partial t} \bar{\mathscr{F}}_t(\alpha, \boldsymbol{n}) \, \mathrm{d}t &= [\mathrm{e}^{-\beta t} \bar{\mathscr{F}}_t(\alpha, \boldsymbol{n})]_0^\infty + \beta \int_{(0,\infty)} \mathrm{e}^{-\beta t} \bar{\mathscr{F}}_t(\alpha, \boldsymbol{n}) \, \mathrm{d}t \\
&= -\mathbf{1}_{\{\boldsymbol{n} = \boldsymbol{n}_0\}} + \beta \, \mathscr{G}_{\boldsymbol{n}}(\alpha, \beta),
\end{aligned}
$$

where we used $\lim_{t \downarrow 0} \bar{\mathscr{F}}_t(\alpha, \boldsymbol{n}) = 0$ for all $\boldsymbol{n} \neq \boldsymbol{n}_0$ and $\lim_{t \downarrow 0} \bar{\mathscr{F}}_t(\alpha, \boldsymbol{n}_0) = P_0(\boldsymbol{n}_0) = 1$. Upon combining the above, we arrive at the desired result. □

Acknowledgements

The first author thanks Onno Boxma (Eindhoven University of Technology), Harsha Honnappa (Purdue University), and Liron Ravner (University of Haifa) for helpful discussions. Both authors thank Werner Scheinhardt (University of Twente) for instructive remarks.

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press

Funding information

The first author's research has been funded by the NWO Gravitation project NETWORKS, grant number 024.002.003.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] ABATE, J. AND WHITT, W. (1995). Numerical inversion of Laplace transforms of probability distributions. *ORSA J. Comput.* **7**, 36–43.

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

[3] BET, G. (2020). An alternative approach to heavy-traffic limits for finite-pool queues. *Queueing Systems* **95**, 121–144.

[4] BET, G., VAN DER HOFSTAD, R. AND VAN LEEUWAARDEN, J. (2019). Heavy-traffic analysis through uniform acceleration of queues with diminishing populations. *Math. Operat. Research* **44**, 821–864.

[5] D'AURIA, B., IVANOVS, J., KELLA, O. AND MANDJES, M. (2010). First passage of a Markov additive process and generalized Jordan chains. *J. Appl. Prob.* **47**, 1048–1057.

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

[7] GLAZER, A. AND HASSIN, R. (1983). ?/M/1: on the equilibrium distribution of customer arrivals. *Europ. J. Operat. Res.* **13**, 146–150.

[8] HAIGHT, F. (1957). Queueing with balking. *Biometrika* **44**, 360–369.

[9] HAVIV, M. AND RAVNER, L. (2021). A survey of queueing systems with strategic timing of arrivals. *Queueing Systems* **99**, 163–198.

[10] HONNAPPA, H. (2017). Rare events of transitory queues. *J. Appl. Prob.* **54**, 943–962.

[11] HONNAPPA, H., JAIN, R. AND WARD, A. (2015). The $\Delta^{(i)}$/GI/1 queueing model, and its fluid and diffusion approximations. *Queueing Systems* **80**, 71–103.

[12] DEN ISEGER, P. (2006). Numerical transform inversion using Gaussian quadrature. *Prob. Eng. Inf. Sci.* **20**, 1–44.

[13] IVANOVS, J., BOXMA, O. AND MANDJES, M. (2010). Singularities of the matrix exponent of a Markov additive process with one-sided jumps. *Stoch. Process. Appl.* **120**, 1776–1794.

[14] JUNEJA, S. AND SHIMKIN, N. (2013). The concert queueing game: strategic arrivals with waiting and tardiness costs. *Queueing Systems* **74**, 369–402.

[15] VAN KREVELD, L., MANDJES, M. AND DORSMAN, J.-P. (2022). Extreme value analysis for a Markov additive process driven by a nonirreducible background chain. *Stoch. Systems* **12**, 293–317.

[16] KUIPER, A., MANDJES, M., DE MAST, J. AND BROKKELKAMP, R. (2023). A flexible and optimal approach for appointment scheduling in healthcare. *Decision Sciences* **54**, 85–100.

[17] MAILLARDET, R. AND TAYLOR, P. (2016). Queues with advanced reservations: an infinite-server proxy for the bookings diary. *Adv. Appl. Prob.* **48**, 13–31.

[18] TIJMS, H. (1986). *Stochastic Modelling and Analysis: A Computational Approach*. Wiley, Chichester.

https://doi.org/10.1017/jpr.2024.77 Published online by Cambridge University Press