FINITE CUSTOMER-POOL QUEUES

ONNO BOXMA, OFFER KELLA, AND MICHEL MANDJES

ABSTRACT. In this paper we consider an M/G/1-type queue fed by a finite customer-pool. In terms of transforms, we characterize the time-dependent distribution of the number of customers and the workload, as well as the associated waiting times.

KEYWORDS. Queueing ∘ finite-customer models ∘ Laplace-Stieltjes transform

AFFILIATIONS. OB is with EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, the Netherlands. OK is with Department of Statistics and Data Science, The Hebrew University of Jerusalem; Jerusalem 9190501, Israel. MM is with the Mathematical Institute, Leiden University, The Netherlands. He is also affiliated with Korteweg-de Vries Institute for Mathematics, University of Amsterdam, The Netherlands; EURANDOM, Eindhoven University of Technology, The Netherlands; Amsterdam Business School, University of Amsterdam, The Netherlands.

ACKNOWLEDGMENTS. The research of OB and MM was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 945045, and by the NWO Gravitation project NETWORKS under grant agreement no. 024.002.003. The research of OK is supported in part by the NSF grant 3336/24 and by the Vigevani Chair in Statistics. Date: *October 7, 2024*.

EMAIL. o.j.boxma@tue.nl, offer.kella@gmail.com, m.r.h.mandjes@math.leidenuniv.nl

1. INTRODUCTION

The model that we consider in this paper can be interpreted as a finite customer-pool M/G/1 system. It has the distinguishing feature that, after time 0, only a finite number m of customers bring work into the system, while there are an additional k customers present at time 0 (whose service has not started at time 0). Once a customer's work has been processed, it leaves the system, entailing that the queue eventually empties with probability 1 (i.e., without any stability requirement being imposed). The model is of M/G/1 type in that in our setup the interarrival times are independent and exponentially distributed with a parameter that depends on the number of customers yet to arrive.

The objective of this study is to describe, for the model introduced in the previous paragraph, the transient behavior of the number of customers present and the workload. Indeed, we succeed in finding a recursive algorithm (of low complexity) that provides us with the joint transform of these two random quantities at an exponentially distributed time. Our findings allow the evaluation of the joint probability mass function (for the number of customers) and density (for the workload) at any given point in time, relying on e.g. the well-developed machinery of [1, 11].

Our work directly relates to two branches of work. (i) In the first place, there is a connection to papers in which the *arrival* times (rather than *interarrival* times) are independent, identically distributed (i.i.d.) random variables. Indeed, when taking in our model the parameter of the exponentially distributed interarrival times proportional to the number of customers yet to arrive, we can equivalently say that the arrival times are i.i.d. exponentially distributed random variables. This setup with independent arrival times is particularly appropriate in instances that the number

of potential customers is relatively low, so that the usual Poisson arrival assumption cannot be justified. The resulting model has been analyzed in [3, 4, 10], with emphasis on scaling limits in the diffusion regime; we in addition refer to the survey [9]. The recent study [14] provides an exact analysis of the workload at an exponentially distributed time, also covering various model variants (additional Poisson arrival stream, balking customers, etc.). (ii) Our model can also be seen as a special case of a reflected Markov additive process [2, Ch. XI] in which the underlying background process is transient. More concretely, the state of the background process represents the number of customers yet to arrive, a process that decreases over time. The queueing model driven by a MAP with a transient background process has been analyzed in [12]; see also [8] for its ruin-theoretic counterpart. In the present paper we show for our specific model, in which the background process is of pure-death type, that the expressions of [12] simplify considerably.

We conclude this introduction by detailing our contributions and the organization of the paper. In Section 2 we present the model and define the key objects of interest. Our first main result, in Section 3, concerns the distribution of the number of customers at an exponentially distributed time (which, by applying inversion, gives us the distribution at any deterministic time). It is also pointed out how the result can be extended to the joint distribution with the workload. Our findings refute the claim in [10] that the finite customer-pool M/G/1 system be 'not amenable to exact analysis', but the obvious price we pay is that we have to settle for results in the transform domain. Then Section 4 presents an analysis of the waiting-time distribution, including tail asymptotics in case the customers' service times are regularly varying. Section 5 presents several ramifications.

2. Model, notation, and objective

In this section we formally introduce our model, state our objectives, and define two sequences of probabilities that play a crucial role throughout our analysis. We throughout denote $a^+ = \max(a, 0)$, $a^- = \min(a, 0)$, LST abbreviates *Laplace Stieltjes Tranform*, 'a.s.' abbreviates *almost surely*, B is some a.s. positive (and finite) random variable with LST $\beta(\alpha) = \mathbb{E}\, \mathrm{e}^{-\alpha B}$ for $\alpha \geqslant 0$. We follow the convention that $\mathbb{N}$ and $\mathbb{N}_0$ denote the sets of, respectively, positive and nonnegative integers.

In our model $k \in \mathbb{N}_0$ customers are already present at time 0, but their service times have not started at time 0. In addition, $m \in \mathbb{N}_0$ customers are to arrive after time 0. With $\mathrm{Exp}(\lambda)$ for $\lambda > 0$ denoting an exponentially distributed random variable with mean λ^{-1}, the random variable $T_m \sim \mathrm{Exp}(\lambda_m)$ is the time until the first arrival, $T_{m-1} \sim \mathrm{Exp}(\lambda_{m-1})$ is the time between the first and second arrivals, etc. All $k + m$ service times are independent (and independent of the arrival process), and distributed like B. Observe that this model can be seen as a finite customer-pool M/G/1 queue.

In the sequel, $\mathbb{E}_{k,m}[\cdot]$ denotes the expectation when starting with $k \in \mathbb{N}_0$ customers being present, and $m \in \mathbb{N}_0$ customers to arrive according to the mechanism described above. The main quantity of interest is $Z(t)$, the number of customers present at time $t \geqslant 0$. Let $T \sim \mathrm{Exp}(\gamma)$, independently of everything else. It is our objective to compute the generating function

$$\mu_{km}(z) = \mathbb{E}_{k,m}[z^{Z(T)}], \quad z \in (0, 1], \tag{1}$$

of the number of customers present at the exponentially distributed time T. Evidently,

$$\mu_{km}(z) = \int_0^\infty \gamma \mathrm{e}^{-\gamma t}\, \mathbb{E}_{k,m}[z^{Z(t)}]\, \mathrm{d}t, \tag{2}$$

so that one can numerically obtain $\mathbb{E}_{k,m}[z^{Z(t)}]$ by applying Laplace inversion; see for instance the fast and accurate procedures presented in [1, 11]. By repeated differentiation and inserting 0, one can obtain the probability distribution of $Z(t)$, while repeated differentiation and inserting 1 yields all (factorial) moments.

In our analysis two sequences of auxiliary probabilities play a key role. To this end, we denote $S_0 = 0$ and $S_n := \sum_{i=1}^{n} T_i$ for any $n \in \{1, \ldots, m\}$.

- The first sequence of probabilities is defined as follows. Let $u_{00} := \mathbb{P}(T > B) = \beta(\gamma)$ and for $n \in \{0, \ldots, m\}$,

$$\begin{aligned} u_{ni} :&= \begin{cases} \mathbb{P}(S_n - S_{n-i} \leqslant B < S_n - S_{n-i-1}, T > B), & i \in \{0, \ldots, n-1\}, \\ \mathbb{P}(S_n \leqslant B, T > B), & i = n, \end{cases} \\ &= \begin{cases} \mathbb{E}\left(\mathrm{e}^{-\gamma B}\, 1_{\{S_n - S_{n-i} \leqslant B < S_n - S_{n-i-1}\}}\right), & i \in \{0, \ldots, n-1\}, \\ \mathbb{E}\left(\mathrm{e}^{-\gamma B}\, 1_{\{S_n \leqslant B\}}\right), & i = n, \end{cases} \end{aligned} \tag{3}$$

  noting that $\sum_{i=0}^{n} u_{ni} = \mathbb{P}(T > B) = \beta(\gamma)$. The object u_{ni} can be interpreted as the probability that there are i arrivals during a single service time, starting with $n \in \{0, \ldots, m\}$ customers who are yet to arrive, jointly with the event $\{T > B\}$.
- The second sequence of probabilities is defined similarly: we let $v_{00} := \mathbb{P}(T \leqslant B) = 1 - \beta(\gamma)$, while for $n \in \{0, \ldots, m\}$,

$$v_{ni} := \begin{cases} \mathbb{P}(S_n - S_{n-i} \leqslant T < S_n - S_{n-i-1}, T \leqslant B), & i \in \{0, \ldots, n-1\}, \\ \mathbb{P}(S_n \leqslant T, T \leqslant B), & i = n, \end{cases} \tag{4}$$

  noting that $\sum_{i=0}^{n} v_{ni} = \mathbb{P}(T \leqslant B) = 1 - \beta(\gamma)$. In this case, v_{ni} can be interpreted as the probability that there are i arrivals until the exponential time T, jointly with $T \leqslant B$. We observe that for $i \in \{0, \ldots, n-1\}$ we can write

$$v_{ni} = \mathbb{P}(S_n - S_{n-i} \leqslant T < S_n - S_{n-i-1}) - \mathbb{P}(S_n - S_{n-i} \leqslant T < S_n - S_{n-i-1}, T > B) \tag{5}$$

  where

$$\mathbb{P}(S_n - S_{n-i} \leqslant T < S_n - S_{n-i-1}) = \frac{\gamma}{\lambda_{n-i} + \gamma} \prod_{j=n-i+1}^{n} \frac{\lambda_j}{\lambda_j + \gamma} \tag{6}$$

  (with an empty product defined to be 1) and, by the memoryless property,

$$\mathbb{P}(S_n - S_{n-i} \leqslant T < S_n - S_{n-i-1}, T > B) = \mathbb{E}[\mathrm{e}^{-\gamma B} 1_{\{S_n - S_{n-i} \leqslant B+T < S_n - S_{n-i-1}\}}]\,. \tag{7}$$

  Along the same lines,

$$v_{nn} = \prod_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \gamma} - \mathbb{E}[\mathrm{e}^{-\gamma B} 1_{\{S_n \leqslant B+T\}}]\,. \tag{8}$$

Remark 1. We note that when $B \sim \mathrm{Exp}(\mu)$ then $B \wedge T$ and $\{T \leqslant B\}$ (as well as $B \wedge T$ and $\{B > T\}$) are independent, with $B \wedge T \sim \mathrm{Exp}(\gamma + \mu)$ and $\mathbb{P}(B \leqslant T) = 1 - \mathbb{P}(B > T) = \frac{\mu}{\gamma+\mu}$. Therefore,

$$\begin{aligned} \frac{u_{ni}}{\mu} = \frac{v_{ni}}{\gamma} &= \frac{1}{\gamma + \mu}\, \mathbb{P}\left(S_n - S_{n-i} \leqslant B \wedge T < S_n - S_{n-i-1}\right) \\ &= \frac{1}{\gamma + \mu}\, \frac{\gamma + \mu}{\lambda_{n-i} + \gamma + \mu} \prod_{j=n-i+1}^{n} \frac{\lambda_j}{\lambda_j + \gamma + \mu} \end{aligned} \tag{9}$$

$$= \frac{1}{\lambda_{n-i} + \gamma + \mu} \prod_{j=n-i+1}^{n} \frac{\lambda_j}{\lambda_j + \gamma + \mu}, \tag{10}$$

defining $\lambda_0 = 0$. ◇

In Section 3 the probabilities u_{ni} and v_{ni} are, in principle, considered known. Observe in particular that they are functions of the parameter γ corresponding to the 'killing time' T. In Section 5 we consider special instances in which u_{ni} and v_{ni} allow closed-form expressions.

3. Number of customers at killing

In the analysis of the conventional M/G/1 queue, one typically works with the embedded process that records the number of customers present at departure epochs. This process is a Markov chain: if it has value $\ell \in \mathbb{N}$, then at the next epoch it is $\ell - 1$ increased by the number of Poisson arrivals during the service time of the served customer.

In our model we exploit a similar idea. Observe that if the number of customers at a certain departure epoch is $\ell \in \mathbb{N}$ and $T > B$, then the number of customers at the next departure epoch will be given by $\ell - 1$ plus the number of arrivals during the corresponding service time (say i). In addition, supposing that there were still $n \in \{0, \ldots, m\}$ customers yet to arrive, this number becomes $n - i$. Therefore, we have from the memoryless property for T and the customers that have not arrived until time B, for any $\ell \in \{1, \ldots, k\}$ that

$$\mathbb{E}_{\ell,n}[z^{Z(T)} 1_{\{T>B\}}] = \sum_{i=0}^{n} \mu_{\ell+i-1,n-i}(z)\, u_{n,i}\,. \tag{11}$$

We proceed by distinguishing various cases. As we point out, the resulting expressions allow a recursive algorithm by which $\mu_{km}(z)$ can be evaluated.

- When $\ell \in \mathbb{N}$ and $T \leqslant B$, then the total number of customers in the system at time T is ℓ plus the number of customers that have arrived by time T. Therefore in this case, for any $n \in \{0, \ldots, m\}$,

$$\mathbb{E}_{\ell,n}[z^{Z(T)} 1_{\{T\leqslant B\}}] = z^{\ell} \sum_{i=0}^{n} z^i\, v_{ni}\,. \tag{12}$$

  Upon combining the displays (11) and (12), this gives the relation, for any $\ell \in \{1, \ldots, k\}$ and $n \in \{0, \ldots, m\}$,

$$\mu_{\ell n}(z) = \sum_{i=0}^{n} \left(\mu_{\ell+i-1,n-i}(z)\, u_{ni} + z^{\ell+i}\, v_{ni} \right)\,. \tag{13}$$

- Supposing that $\ell = 0$, $n \in \{1, \ldots, m\}$ and $T \leqslant T_n$, we evidently have that $Z(T) = 0$. In case $\ell = 0$, $n \in \{1, \ldots, m\}$ and $T > T_n$, then by the memoryless property, $Z(T)$ behaves as if we would have started from the state $(1, n-1)$. As a result,

$$\mu_{0n}(z) = \frac{\gamma}{\gamma + \lambda_n} + \frac{\lambda_n}{\gamma + \lambda_n}\, \mu_{1,n-1}(z)\,. \tag{14}$$

- Finally, when starting with $\ell = n = 0$, then clearly $Z(T) = 0$ and thus $\mu_{00}(z) = 1$. For $n = 0$ and $\ell \in \{1, \ldots, k\}$, (13) becomes

$$\mu_{\ell 0}(z) = \mu_{\ell-1,0}(z)\, u_{00} + z^{\ell}\, v_{00}\,. \tag{15}$$

This implies, by induction, that

$$\mu_{\ell 0}(z) = \begin{cases} (1 - v_{00})u_{00}^{\ell} + v_{00}\,\dfrac{u_{00}^{\ell+1} - z^{\ell+1}}{u_{00} - z}, & z \neq u_{00}, \\ (1 + v_{00}\ell)\,u_{00}^{\ell}, & z = u_{00}\,. \end{cases} \tag{16}$$

The above can be used to devise a procedure that recursively determines $\mu_{km}(z)$. It is given by the following algorithm.

ALGORITHM to evaluate $\mu_{km}(z)$.
INPUT: k, m, u_{ni} and v_{ni} for all $n \in \{0,\dots,m\}$ and $i \in \{0,\dots,n\}$.
OUTPUT: $\mu_{\ell n}(z)$ for all $\ell \in \{0,\dots,k\}$ and $n \in \{0,\dots,m\}$.

01: $\mu_{00}(z) \leftarrow 1$;
02: FOR $\ell = 1$ TO k DO
03: Compute $\mu_{\ell 0}(z)$ via (16);
04: END;
05: FOR $n = 1$ TO m DO
06: Compute $\mu_{0n}(z)$ via (14);
07: FOR $\ell = 1$ TO k DO
08: Compute $\mu_{\ell n}(z)$ via (13);
09: END;
10: END;
11: RETURN $\mu_{km}(z)$.

Observe that in each step of this algorithm, one only needs objects that have been computed earlier.

Theorem 1. *The probability generating function $\mu_{km}(z)$ can be recursively identified via the above algorithm. If u_{ni} and v_{ni}, for all $n \in \{0,\dots,m\}$ and $i \in \{0,\dots,n\}$, are known, then the complexity of the algorithm is $O(km^2)$.*

The complexity of this algorithm is $O(km^2)$ because the complexity of the for-loop in lines 07–09 is $O(kn)$, and this loop has to be performed for $n = 1$ up to m.

Remark 2. One can easily adapt our algorithm to facilitate the computation of the joint transform of the number of customers and the work in the system at the exponential time T. Namely, with mild abuse of notations, we first replace $\mu_{\ell n}(z)$ by

$$\mu_{\ell n}(z, \alpha) := \mathbb{E}_{\ell n}[z^{Z(T)}\mathrm{e}^{-\alpha W(T)}], \tag{17}$$

where $W(t)$ is the amount of work at time t. Note that the quantity u_{ni} remains unchanged, but the v_{ni} needs to be adapted. To this end we observe that, starting from state (ℓ, n), with $\ell \in \{1,\dots,k\}$, then when $T \leqslant B$ and the number of arrivals until time T is i, then the total number in the system at time T is $\ell + i$ and the total amount of work at time T is $B - T$ plus a sum of $\ell + i - 1$ independent random variables (which are also independent of everything else) that are distributed like the generic service time B. Therefore, we see that $z^{\ell+i}v_{ni}$ in (13) needs to be replaced by

$$z^{\ell+i}\beta(\alpha)^{\ell+i-1}v_{ni}(\alpha) \tag{18}$$

where $v_{00}(\alpha) = \mathbb{E}[\mathrm{e}^{-\alpha(B-T)}1_{\{T\leqslant B\}}]$ and, for $n \in \{1,\dots,m\}$,

$$v_{ni}(\alpha) = \begin{cases} \mathbb{E}[\mathrm{e}^{-\alpha(B-T)}1_{\{S_n - S_{n-i} \leqslant T < S_n - S_{n-i-1}, T\leqslant B\}}], & i \in \{0,\dots,n-1\}, \\ \mathbb{E}[\mathrm{e}^{-\alpha(B-T)}1_{\{S_n \leqslant T, T\leqslant B\}}], & i = n. \end{cases} \tag{19}$$

As before, we have that $\mu_{00}(z,\alpha) = 1$, and (14), (15) and (16) remain unchanged, apart from the obvious replacement of $\mu_{\ell n}(z)$ by $\mu_{\ell n}(z,\alpha)$, and of v_{00} by $v_{00}(\alpha)$. In particular, the structure of the algorithm remains the same.
Clearly, setting $z = 1$ (and then $z^{\ell+i} = 1$ in the only place where this is relevant), immediately gives an algorithm for the computation of the LST of the workload in the system at an independent exponentially distributed time, starting from the state (k,m). ◇

Remark 3. So far we have assumed that the service times are identically distributed and that the interarrival times $T_m,\ldots,T_1$ are exponentially distributed with parameters $\lambda_m,\ldots,\lambda_1$. This can be substantially generalized, in that we could instead assume that, when after a service completion (or at time zero) there are n customers to arrive and ℓ customers present, the next service time is distributed like some $B_{\ell n}$ and that the interarrival times (until the next service completion) have parameters $\lambda_{\ell n},\ldots,\lambda_{\ell 1}$, while maintaining all the underlying independence assumptions.
For this generalization, the only change that is needed in the description is that u_{ni} and v_{ni} are replaced by u_{kni} and v_{kni} where the latter are defined via (3) and (4) in which B is replaced by B_{kn} and S_i are similarly defined only with $T_i \sim \mathrm{Exp}(\lambda_{ki})$, $i = 1,\ldots,m$. All the rest of the development is identical. ◇

Remark 4. We observe that, by differentiation of (13), (14) and (15) (or (16) directly) ℓ times and letting $z \uparrow 1$, then exactly the same algorithm results in the recursive computation of the factorial moments

$$\mathbb{E}_{km}(Z(T)(Z(T)-1)\ldots(Z(T)-\ell+1)) \tag{20}$$

for any desired k,m,ℓ, from which the moments $\mathbb{E}_{km}Z^{\ell}(T)$ may be deduced in the usual way.
Similarly, equating the coefficients of the polynomials on the two hand sides of (13), (14) and (15), results in a similar algorithm for the recursive computation of the probabilities $\mathbb{P}_{km}(Z(T) = \ell)$. This is a straightforward exercise. ◇

Remark 5. In this remark we study a process that is useful in the context of the waiting-time analysis of Section 4. We consider the bivariate Markov chain $(Z_h,V_h)_h$, with $h = 1,\ldots,k+m$, where (i) Z_h denotes the number of customers that is in the system after the h-th departure, and (ii) V_h denotes the number of customers, out of the m that are arriving after time 0, that still need to arrive after the h-th departure. Note that V_h are non-increasing in h, almost surely, and that $(Z_0,V_0) = (k,m)$. In order to analyze the Markov chain $(Z_h,V_h)_h$, we introduce the following sequence of probabilities:

$$w_{ni} := \begin{cases} \mathbb{P}(S_n - S_{n-i} \leqslant B < S_n - S_{n-i-1}), & i \in \{0,\ldots,n-1\}, \\ \mathbb{P}(S_n \leqslant B), & i = n. \end{cases} \tag{21}$$

Observe that in the limiting regime that $\gamma \downarrow 0$, the probabilities w_{ni} and u_{ni} coincide.
The transition probabilities of $(Z_h,V_h)_h$ are given by

$$p_{(\ell_1,n_1),(\ell_2,n_2)} = \begin{cases} w_{n_1,n_1-n_2}, & \ell_1 \in \{1,\ldots,k\},\ n_2 \in \{0,\ldots,n_1\},\ \ell_2 = \ell_1+n_1-n_2-1, \\ w_{n_1-1,n_1-n_2}, & \ell_1 = 0,\ n_1 \in \{1,\ldots,m\},\ n_2 \in \{0,\ldots,n_1-1\}, \ell_2 = n_1-n_2-1, \\ 1, & \ell_1 = n_1 = \ell_2 = n_2 = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

These cases can be interpreted as follows. The first case corresponds to the scenario that the queue is non-empty at the end of the previous service time, then a new service time starts, during this

service time $n_1 - n_2$ customers arrive, with n_1 customers that are yet to arrive at the beginning of the service time. The second case corresponds to a similar scenario, but now the queue is empty at the end of the previous service time, and a new service time starts as soon as a new customer enters (so that $n_1 - 1$ customer arrive during this service time). The third case corresponds to the absorbing state $(0, 0)$.

As for u_{ni} and v_{ni} (see Section 5), it is easy to check that w_{ni} also have an explicit form for the cases where either $\lambda_i = \lambda$ for $i = 1, \ldots, n$ or $\lambda_i = i\lambda$ for $i = 1, \ldots, n$. These transition probabilities will be needed in the next section; more specifically, they are needed to determine the probability that the system is empty after a service completion. ◇

4. Waiting times

As before, we consider the setting in which k customers are already present at time 0, while in addition m customers are yet to arrive after time 0, with the corresponding interarrival times being represented by $T_m, T_{m-1}, \ldots, T_1$, where $T_i \sim \text{Exp}(\lambda_i)$ for $i = 1, \ldots, m$. We throughout assume a FIFO service discipline, where the initial k customers are arranged at some arbitrary order. With $W_1, \ldots, W_k$ denoting the waiting times of those customers already present at zero, and measuring their waiting times from time 0 on, we obviously have

$$\mathbb{E}[e^{-\alpha W_j}] = \beta^{j-1}(\alpha), \quad j = 1, \ldots, k. \tag{23}$$

The goal of this section is therefore to identify, in terms of transforms, the waiting time distributions of the m customers arriving after time 0, denoted by $W_{k+1}, \ldots, W_{k+m}$.

The waiting times of the customers arriving after time 0 obey the Lindley recursion

$$W_{j+1} = [W_j + B_j - I_j]^+, \quad j = k, \ldots, k + m - 1, \tag{24}$$

where I_j denotes the interarrival time between arriving customers $j - k$ and $j - k + 1$ (for $j = k$: the arrival time of customer $k + 1$). We can use the identity $e^{-\alpha x^+} = e^{-\alpha x} - e^{-\alpha x^-} + 1$, to conclude that, for $i = 0, \ldots, m - 1$,

$$\mathbb{E}[e^{-\alpha W_{k+i+1}}] = \mathbb{E}[e^{-\alpha [W_{k+i}+B_{k+i}-I_{k+i}]^+}] = \mathbb{E}[e^{-\alpha (W_{k+i}+B_{k+i}-I_{k+i})}] - \mathbb{E}[e^{-\alpha [W_{k+i}+B_{k+i}-I_{k+i}]^-}] + 1. \tag{25}$$

Using the independence of the components in the first term in the righthand side, and also using the fact that $I_{k+i} \sim \text{Exp}(\lambda_{m-i})$ (implying that it is memoryless), we obtain from (25) that

$$\begin{aligned} \mathbb{E}[e^{-\alpha W_{k+i+1}}] &= \mathbb{E}[e^{-\alpha W_{k+i}}]\beta(\alpha)\frac{\lambda_{m-i}}{\lambda_{m-i} - \alpha} - \mathbb{P}(W_{k+i+1} = 0)\frac{\lambda_{m-i}}{\lambda_{m-i} - \alpha} - \mathbb{P}(W_{k+i+1} > 0) + 1 \\ &= \mathbb{E}[e^{-\alpha W_{k+i}}]\beta(\alpha)\frac{\lambda_{m-i}}{\lambda_{m-i} - \alpha} - \mathbb{P}(W_{k+i+1} = 0)\frac{\alpha}{\lambda_{m-i} - \alpha}. \end{aligned} \tag{26}$$

Iterating this relation, we readily obtain the following result. In the sequel we use the notation $\varrho_h := \mathbb{P}(W_h = 0)$, for $h = k + 1, \ldots, k + m$.

Theorem 2. *For $j = k + 1, \ldots, k + m$, empty products being equal to one,*

$$\mathbb{E}[e^{-\alpha W_j}] = \beta^{j-1}(\alpha) \prod_{i=0}^{j-k-1} \frac{\lambda_{m-i}}{\lambda_{m-i} - \alpha} - \sum_{h=k+1}^{j} \varrho_h \beta^{j-h}(\alpha) \frac{\alpha}{\lambda_{m-h+k+1} - \alpha} \prod_{w=0}^{j-1-h} \frac{\lambda_{m+k-h-w}}{\lambda_{m+k-h-w} - \alpha}. \tag{27}$$

We have thus derived an expression for the waiting time LST of all $k + m$ customers, in known quantities and in the constants ϱ_h for $h = k + 1, \ldots, k + m$. Now observe that $\varrho_h = \mathbb{P}(Z_{h-1} = 0)$. Hence one only needs to determine the latter probabilities from the one-step transition probabilities $\mathbb{P}(Z_{i+1} = j \mid Z_i = r)$ which are discussed in Remark 5 above.

Differentiating (27) with respect to α and subsequently taking $\alpha = 0$, we obtain that if $\mathbb{E}[B] < \infty$, then, for $j = k+1, \ldots, k+m$:

$$\mathbb{E}[W_j] = (j-1)\mathbb{E}[B] - \sum_{i=0}^{j-k-1} \frac{1}{\lambda_{m-i}} + \sum_{h=k+1}^{j} \varrho_h \frac{1}{\lambda_{m-h+k+1}}. \tag{28}$$

Conclude in particular that if $\mathbb{E}[B] < \infty$, then $\mathbb{E}[W_j] < \infty$.

We end this section by exploiting (27) and (28) to derive the tail asymptotics of the distribution of waiting time W_j in the case of a regularly varying service-time distribution, an important class of heavy-tailed distributions. We use the well-known Bingham-Doney lemma, cf. [5, Theorem 8.1.6], which states that (i) and (ii) below are equivalent: (i) B is regularly varying of index $-\nu \in (-2,-1)$ at infinity, i.e.,

$$\mathbb{P}(B > t) \sim \frac{-1}{\Gamma(1-\nu)} t^{-\nu} L(t), \quad t \to \infty, \tag{29}$$

with $L(t)$ a slowly-varying function (entailing that the first moment of B is finite but the second is not), and (ii)

$$\beta(\alpha) - 1 + \alpha\mathbb{E}[B] \sim \alpha^{\nu} L(1/\alpha), \quad \alpha \downarrow 0. \tag{30}$$

Now assume that (29) and hence also (30) hold. It then readily follows from (27) and (28) that

$$\mathbb{E}[\mathrm{e}^{-\alpha W_j}] - 1 + \alpha\mathbb{E}[W_j] \sim (j-1)\alpha^{\nu} L(1/\alpha), \quad \alpha \downarrow 0. \tag{31}$$

Once more applying the equivalence between tail behaviour of the distribution and behavior of the LST near zero that is specified by the Bingham-Doney lemma, but now for W_j, we find the following result. We denote by $f(t) \sim g(t)$ that $f(t)/g(t) \to 1$ as $t \to \infty$.

Proposition 1. *Suppose the service-time distribution obeys (29). Then,*

$$\mathbb{P}(W_j > t) \sim (j-1)\mathbb{P}(B > t), \quad t \to \infty. \tag{32}$$

For the ordinary M/G/1 queue in steady state, with regularly varying service time distribution as in (29), it is well-known [6] that the waiting time is also regularly varying but with index $1-\nu$. However, in our setting with only a finite number of customers, the most likely scenario for a very large waiting time W_j is that one of the preceding $j-1$ service times is very large, resulting in (32); in this respect, bear in mind that for independent $B_1, \ldots, B_j$ that are distributed as the random variable B, we have that

$$\mathbb{P}\left(\sum_{i=1}^{j} B_i > t\right) \sim j\,\mathbb{P}(B > t), \quad t \to \infty. \tag{33}$$

5. Special cases

We now consider two special cases in which the required probability sequences u_{mi} and v_{mi} allow an explicit evaluation. In the first case we take $\lambda_i = i\lambda$. This means that the random vector $(T_m, \ldots, T_1)$ is distributed as the differences of the components of the order statistics associated with m i.i.d. Exp(λ) distributed random variables. This arrival process is the pure death process analogue of the Yule process (which is a pure birth process). We can treat it as the case where the m customers who are yet to arrive pick their arrival times according to independent (identically distributed) exponential clocks. Whereas for the Poisson process the inter-arrival times are i.i.d. exponentially

distributed, for the case with $\lambda_i = i\lambda$ the arrival times are i.i.d. exponentially distributed. Hence, as is easily checked, for $i \in \{0, \ldots, m\}$,

$$u_{mi} = \binom{m}{i} \mathbb{E}\left(\mathrm{e}^{-\gamma B}(1-\mathrm{e}^{-\lambda B})^i \, \mathrm{e}^{-(m-i)\lambda B}\right), \tag{34}$$

To compute v_{mi} we again recall (4), (5), (6), (7) and (8), so as to obtain, for $i \in \{0, \ldots, m\}$,

$$v_{mi} = \frac{\gamma}{\gamma + (m-i)\lambda} \prod_{j=m-i+1}^{m} \frac{j\lambda}{\gamma + j\lambda} - \binom{m}{i} \mathbb{E}\left(\mathrm{e}^{-\gamma B}(1-\mathrm{e}^{-\lambda(B+T)})^i \, \mathrm{e}^{-(m-i)\lambda(B+T)}\right). \tag{35}$$

The expressions for u_{mi} and v_{mi} can be made more explicit by noting that

$$\mathbb{E}\left(\mathrm{e}^{-\gamma B}(1-\mathrm{e}^{-\lambda B})^i \, \mathrm{e}^{-(m-i)\lambda B}\right) = \sum_{j=0}^{i} \binom{i}{j} (-1)^{i-j} \beta(\gamma + (m-j)\lambda), \tag{36}$$

$$\mathbb{E}\mathrm{e}^{-\gamma B}\left((1-\mathrm{e}^{-\lambda(B+T)})^i \, \mathrm{e}^{-(m-i)\lambda(B+T)}\right) = \sum_{j=0}^{i} \binom{i}{j} (-1)^{i-j} \beta(\gamma + (m-j)\lambda) \, \frac{\gamma}{\gamma + (m-j)\lambda}. \tag{37}$$

The remainder of this section covers a second case in which $\lambda_i \equiv \lambda$, i.e., arrivals occur according to a Poisson process with rate λ which is stopped after the m-th arrival. By Eqn. (3),

$$u_{mi} = \begin{cases} \mathbb{E}\left(\mathrm{e}^{-(\lambda+\gamma)B} \frac{(\lambda B)^i}{i!}\right), & i \in \{0, \ldots, m-1\}, \\ \beta(\gamma) - \sum_{j=0}^{m-1} u_{mj}, & i = m\,. \end{cases} \tag{38}$$

Also, recalling (4), (5), (6), (7) and (8),

$$v_{mi} = \begin{cases} \left(\frac{\lambda}{\lambda+\gamma}\right)^i \frac{\gamma}{\gamma+\lambda} - \mathbb{E}\left(\mathrm{e}^{-((\lambda+\gamma)B+\lambda T)} \frac{(\lambda(B+T))^i}{i!}\right), & i \in \{0, \ldots, m-1\} \\ 1 - \beta(\gamma) - \sum_{j=0}^{m-1} v_{mj}, & i = m\,. \end{cases} \tag{39}$$

Observe that in (38) and (39) we can also write, for $i \in \{0, \ldots, m-1\}$,

$$\mathbb{E}\left(\mathrm{e}^{-(\lambda+\gamma)B} \frac{(\lambda B)^i}{i!}\right) = \frac{(-\lambda)^i \beta^{(i)}(\lambda+\gamma)}{i!} \tag{40}$$

$$\begin{aligned} \mathbb{E}\left(\mathrm{e}^{-((\lambda+\gamma)B+\lambda T)} \frac{(\lambda(B+T))^i}{i!}\right) &= \sum_{j=0}^{i} \mathbb{E}\left(\mathrm{e}^{-(\lambda+\gamma)B} \frac{(\lambda B)^j}{j!}\right) \mathbb{E}\left(\mathrm{e}^{-\lambda T} \frac{(\lambda T)^{i-j}}{(i-j)!}\right) \\ &= \sum_{j=0}^{i} \frac{(-\lambda)^j \beta^{(j)}(\lambda+\gamma)}{j!} \, \frac{\gamma}{\lambda+\gamma} \left(\frac{\lambda}{\lambda+\gamma}\right)^{i-j}. \end{aligned} \tag{41}$$

Throughout this paper we have assumed that k customers are initially present, and there is a pool of m customers who arrive after time zero. Below we briefly study the case in which both k and m are, geometrically distributed, random variables. Restricting ourselves to the case $\lambda_i \equiv \lambda$ and to Exp(μ) distributed service times, we obtain an explicit expression for the generating function of the number of customers at an Exp(γ) distributed time (when divided by γ, this is also the Laplace transform of the generating function of the number of customers at time t). This generalizes a known result (cf. [7, Ch. II.2]) for the transient behavior of the ordinary M/M/1 queue to the case in which the arrival process dries up after a geometrically distributed number of arrivals. In this case, recalling Remark 1 and with $\xi \equiv \xi(\lambda, \mu, \gamma) := \lambda + \mu + \gamma$, (38) and (39) become

$$u_{mi} = \begin{cases} \frac{\mu}{\xi}\left(\frac{\lambda}{\xi}\right)^i, & i \in \{0,1,\ldots,m-1\}, \\ \frac{\mu}{\mu+\gamma}\left(\frac{\lambda}{\xi}\right)^m, & i = m\,, \end{cases} \tag{42}$$

and

$$v_{mi} = \begin{cases} \frac{\gamma}{\xi}\left(\frac{\lambda}{\xi}\right)^i, & i \in \{0,1,\ldots,m-1\}, \\ \frac{\gamma}{\mu+\gamma}\left(\frac{\lambda}{\xi}\right)^m, & i = m\,. \end{cases} \tag{43}$$

The above equations have obvious interpretations. For instance, regarding (42): among three Poisson events with rates λ, μ, γ, an arrival occurs i times first with probability $(\lambda/\xi)^i$, and a service completion comes next with probability μ/ξ (but with probability $\mu/(\mu+\gamma)$ if there is no potential arrival left).

Introduce, for $p, r \in [0,1]$ and $z \in (0,1]$,

$$G(p,r,z) := \sum_{\ell=0}^{\infty} p^\ell \sum_{n=0}^{\infty} r^n \mu_{\ell,n}(z), \qquad M_\ell(r,z) := \sum_{n=0}^{\infty} r^n \mu_{\ell,n}(z), \quad \ell = 0,1. \tag{44}$$

Observe that $G(p,r,z)$ is the generating function of $\mu_{\ell,n}(z)$ in both ℓ and n, but it should be multiplied by $(1-p)(1-r)$ to obtain the generating function of $Z(T)$ for geometrically distributed k and m. Note that, applying numerical inversion to $G(p,r,z)$, we can identify $\mu_{\ell,n}(z)$, and hence we can in principle also deal with arbitrarily distributed k and m. In the rest of this section we find a closed-form expression for $G(p,r,z)$. Taking generating functions in (13) yields:

$$G(p,r,z) - M_0(r,z) = \sum_{\ell=1}^{\infty} p^\ell \sum_{n=0}^{\infty} r^n \sum_{i=0}^{n} \mu_{\ell+i-1,n-i}(z)\, u_{ni} + \sum_{\ell=1}^{\infty} p^\ell \sum_{n=0}^{\infty} r^n \sum_{i=0}^{n} z^{\ell+i}\, v_{ni} =: \mathrm{I} + \mathrm{II}. \tag{45}$$

Recall the identity $(a-b)\sum_{j=0}^{\infty} a^j \sum_{i=0}^{\infty} b^i c_{j+i} = a\sum_{h=0}^{\infty} c_h a^h - b\sum_{h=0}^{\infty} c_h b^h$ for any $a, b \in [0,1]$ and a sequence c_k. Hence,

$$\begin{aligned} G(p,r,z) - M_0(r,z) = {} & \frac{\mu p}{\xi - \lambda r/p} G(p,r,z) - \frac{\lambda r}{\xi}\frac{\mu}{\xi - \lambda r/p} G\left(\frac{\lambda r}{\xi}, r, z\right) + \\ & \frac{p\mu}{\mu+\gamma-p\mu} H_1(r,z) - \frac{pz}{1-pz} H_2(r,z) + \frac{pz}{1-pz} F(r,z), \end{aligned} \tag{46}$$

where

$$H_1(r,z) := \frac{(\mu+\gamma)(1-z)}{\mu - (\mu+\gamma)z} \frac{\lambda\mu}{(\mu+\gamma)\xi - r\mu\lambda}, \tag{47}$$

$$H_2(r,z) := \frac{\gamma}{\mu - (\mu+\gamma)z} \frac{\mu}{\mu+\gamma} \frac{\lambda}{\xi - \lambda r z}, \tag{48}$$

$$F(r,z) := \frac{\gamma}{\mu+\gamma} \frac{1}{1-r} \frac{\xi - \lambda r}{\xi - \lambda r z}. \tag{49}$$

Here the first four terms in the right-hand side of (46) concern I, and the fifth term II. The third and fourth term correct for taking the expression in the first (instead of the second) line of (42) for u_{nn}. Observe that $z = \mu/(\mu+\gamma)$ is a pole of both the third and the fourth term, but *not* of their difference. In the fifth term we have used both parts of (43), and also the first line of (16). When $z = u_{00}$, hence $z = \mu/\xi$, one should use the second line of (16); we have omitted that calculation.

The next step is to rewrite (46) into

$$G(p,r,z)\left(\frac{\mu p^2-\xi p+\lambda r}{\lambda r-\xi p}\right)=M_0(r,z)+\frac{\lambda r}{\xi}\frac{\mu p}{\lambda r-\xi p}G\left(\frac{\lambda r}{\xi},r,z\right)$$
$$+\ \frac{p\mu}{\mu+\gamma-p\mu}H_1(r,z)-\frac{pz}{1-pz}H_2(r,z)+\frac{pz}{1-pz}F(r,z). \quad (50)$$

It is readily verified that the numerator of the term between brackets in the left-hand side of (50) has one zero, say $p^\star$, in $(0,1)$ (use Rouché's theorem, or take $p=0$ and $p=1$ in that numerator), and a second zero $\widehat{p}=\lambda r/(\mu p^\star)>1$. Since $G(p^\star,r,z)$ is finite, the right-hand side of (50) must be zero for $p=p^\star$, yielding a relation between the unknown functions $G(\frac{\lambda r}{\xi},r,z)$ and $M_0(r,z)$: with $J(r,z):=H_2(r,z)-F(r,z)$,

$$\frac{\lambda r}{\xi}\frac{\mu p^\star}{\lambda r-\xi p^\star}G\left(\frac{\lambda r}{\xi},r,z\right)=\ -M_0(r,z)-\frac{p^\star\mu}{\mu+\gamma-p^\star\mu}H_1(r,z)+\frac{p^\star z}{1-p^\star z}J(r,z). \quad (51)$$

Substituting this in (50) gives, after some calculations,

$$\mu(p-\widehat{p})G(p,r,z)=\ -\frac{\lambda r}{p^\star}M_0(r,z)+\frac{p\mu}{\mu+\gamma-p\mu}\frac{\lambda\mu r-(\mu+\gamma)\xi}{(\mu+\gamma-p^\star\mu)}H_1(r,z)-$$
$$\frac{pz}{1-pz}\frac{\lambda rz-\xi}{1-p^\star z}J(r,z). \quad (52)$$

It remains to determine $M_0(r,z)$. Using the one equation we have not used yet, namely (14), in combination with $\mu_{0,0}(z)\equiv 1$, we obtain

$$M_0(r,z)-1=\frac{\gamma}{\lambda+\gamma}\frac{r}{1-r}+\frac{\lambda}{\lambda+\gamma}rM_1(r,z). \quad (53)$$

From (44), in combination with (53), we see that

$$\lim_{p\downarrow 0}\frac{G(p,r,z)-M_0(r,z)}{p}=M_1(r,z)=\frac{\lambda+\gamma}{\lambda r}\left(M_0(r,z)-1-\frac{\gamma}{\lambda+\gamma}\frac{r}{1-r}\right). \quad (54)$$

Subtracting $\mu(p-\widehat{p})M_0(r,z)$ from both sides in (52), dividing by p and finally letting $p\downarrow 0$ results, in combination with (54), in an equation from which $M_0(r,z)$ can be determined (notice that, in combining the resulting two $M_0(r,z)$ terms in the right-hand side, we have used that $p^\star\widehat{p}=\lambda r/\mu$):

$$\frac{\lambda+\gamma}{\lambda r}\left(M_0(r,z)-1-\frac{\gamma}{\lambda+\gamma}\frac{r}{1-r}\right)$$
$$=\frac{M_0(r,z)}{\widehat{p}}-\frac{1}{\mu\widehat{p}}\left(\frac{\mu}{\mu+\gamma}\frac{\lambda\mu r-(\mu+\gamma)\xi}{\mu+\gamma-p^\star\mu}H_1(r,z)-z\frac{\lambda rz-\xi}{1-p^\star z}J(r,z)\right). \quad (55)$$

Hence

$$M_0(r,z)=\frac{(\lambda+\gamma)\widehat{p}}{(\lambda+\gamma)\widehat{p}-\lambda r}+\frac{\gamma r\widehat{p}}{(1-r)((\lambda+\gamma)\widehat{p}-\lambda r}-$$
$$\frac{\lambda r}{\mu}\frac{1}{(\lambda+\gamma)\widehat{p}-\lambda r}\left(\frac{\mu}{\mu+\gamma}\frac{\lambda\mu r-(\mu+\gamma)\xi}{\mu+\gamma-p^\star\mu}H_1(r,z)-z\frac{\lambda rz-\xi}{1-p^\star z}J(r,z)\right). \quad (56)$$

Notice that $M_0(0,z)=1$, as it should be.

In conclusion, we find the following result. Recall that $\xi\equiv\xi(\lambda,\mu,\gamma)=\lambda+\mu+\gamma$.

Proposition 2. *For any $p,r\in[0,1]$ and $z\in(0,1]$, $G(p,r,z)$ is given by* (52)*, with $M_0(r,z)$ given by* (56)*.*

REFERENCES

[1] J. ABATE and W. WHITT (1995). Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing* **7**, pp. 36-43.

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

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

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

[5] N. BINGHAM, C. GOLDIE, and J. TEUGELS (1987). *Regular Variation*. Cambridge University Press, Cambridge.

[6] J.W. COHEN (1973). Some results on regular variation for distributions in queueing and fluctuation theory. *Journal of Applied Probability* **10**, pp. 343-353.

[7] J.W. COHEN (1982). *The Single Server Queue*. North-Holland Publ. Cy., Amsterdam.

[8] G. DELSING and M. MANDJES (2021). A transient Cramér-Lundberg model with applications to credit risk. *Journal of Applied Probability* **58**, pp. 721-745.

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

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

[11] P. DEN ISEGER (2006). Numerical transform inversion using Gaussian quadrature. *Probability in the Engineering and Informational Sciences* **20**, pp. 1-44.

[12] L. VAN KREVELD, M. MANDJES, and J.-P. DORSMAN (2022). Extreme value analysis for a Markov Additive Process driven by a nonirreducible background chain. *Stochastic Systems* **12**, pp. 293-317.

[13] M. MANDJES and O. BOXMA (2023). *The Cramér-Lundberg Model and its Variants – A Queueing Perspective.* Springer, New York.

[14] M. MANDJES and D. RUTGERS ($\geqslant$ 2024). A queue with independent and identically distributed arrivals. *Journal of Applied Probability*, to appear.