The M/G/1 queue with quasi-restricted accessibility

The $M/G/1$ queue with quasi-restricted accessibility

Onno Boxma[*], David Perry[†], Wolfgang Stadje[‡] and Shelley Zacks[§]

Abstract

We consider single-server queues of the $M/G/1$ kind with a special kind of partial customer rejection called quasi-restricted accessibility (QRA). Under QRA, the actual service time assigned to an arriving customer depends on his service requirement, say x, the current workload, say w, and a prespecified threshold b. If $x + w \leq b$ the customer is fully served. If $w \leq b < w + x$, the customer receives service time $b - w + (w + x - b)f$ for some random number $f \in [0, 1]$, while if $w > b$ the actual service time is the fraction fx of the requirement. The random fractions are assumed to be i.i.d. We derive the steady-state distribution of the workload process, which is also the steady-state distribution of the waiting time, and provide explicit results for the cases of Erlang or hyperexponential service requirements and uniformly distributed or constant fractions. We also deal with the case of exponential barriers b (instead of one constant threshold). Furthermore, the distribution functions of the length of a busy period and of the cycle maximum of the workload are determined. In the case of phase-type service requirements there is an alternative (martingale) technique to derive the busy period distribution; we illustrate this approach in the case of Erlang(2,μ). Finally, we show in the example of the Erlang(2,μ)/M/1-type QRA queue with deterministic fractions (which is non-Markovian) how to compute the busy period distribution via a duality with a Markovian system.

1 Introduction

In this paper we consider single-server queues with a special kind of partial customer rejection called quasi-restricted accessibility (QRA). Any queueing system with a restricted workload capacity has to deal with the problem what to do with customers whose acceptance would increase the current workload beyond the given limitations. One possibility is to reject these customers completely; another is to grant them only partial admission in order to keep the capacity bounded or to diminish the growth of the cumulated workload.

QRA means that only a certain fraction of any freshly arriving workload above a certain threshold $b > 0$ is processed. Formally, consider a single-server queue of the $M/G/1$ type with A_n denoting the time between the arrivals of the nth and $(n+1)$st customer and X_n denoting the service requirement of the nth customer, $n = 0, 1, 2, \ldots$ Let $F_n,\ n = 0, 1, 2, \ldots$ be random variables taking values in $[0, 1]$. We assume that $(A_n)_{n\geq 0}$, $(X_n)_{n\geq 0}$ and $(F_n)_{n\geq 0}$

[*] EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, HG 9.14, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (boxma@win.tue.nl)

[†] Department of Statistics, University of Haifa, Haifa 31909 Israel (dperry@haifa.ac.il)

[‡] Department of Mathematics and Computer Science, University of Osnabrück, 49069 Osnabrück, Germany (wolfgang@mathematik.uni-osnabrueck.de)

[§] Binghamton University, Department of Mathematical Sciences, Binghamton, NY 13902-6000, USA (shelly@math.binghamton.edu)

are i.i.d. sequences and independent of each other. Furthermore, A_n is $\exp(\lambda)$-distributed for every $n \geq 1$. In the QRA model the *actual service time* S_n assigned to the nth customer depends on his service requirement X_n and his waiting time W_n in the following recursively defined way: Let $W_0 = 0$. If $W_0, ..., W_n$ have already been defined for some $n \geq 0$, let

$$S_n = \begin{cases} X_n, & \text{if } W_n + X_n \leq b \\ b - W_n + F_n(W_n + X_n - b), & \text{if } W_n \leq b,\ W_n + X_n > b \\ F_n X_n, & \text{if } W_n > b \end{cases} \tag{1}$$

and $W_{n+1} = \max[0, W_n + S_n - A_n]$.

Thus if $W_n + X_n$ exceeds b, then only the fraction F_n of the service requirement beyond b is added to the workload. We will deal in particular with the case of deterministic $F_n \equiv f \in [0, 1)$ and the case when F_n is uniform on $(0, 1)$. Note that $F_n \equiv 0$ means that the system does not tolerate more than b work units; arriving jobs whose admission would increase the current workload above b are only partially admitted such that the workload jumps up to its capacity bound b. If $F_n \equiv f \in (0, 1)$, only the fraction f of any workload exceeding b or arriving while the current workload is larger than b is processed. In general, workload above b is either rejected (with probability $\mathbb{P}(F_n = 0)$) or only the fraction F_n is processed if $F_n > 0$.

We denote by A, X, F generic independent random variables whose distributions are those of A_n, X_n, F_n, respectively. Let $\mathbf{V} = \{V(t) \mid t \geq 0\}$ be the associated virtual waiting time process. The system is stable under the condition

$$\lambda \mathbb{E}[F]\mathbb{E}[X] < 1. \tag{2}$$

We assume throughout that (2) holds. Then the waiting times W_n possess a limiting distribution which, by PASTA, is also the limiting (and steady-state) distribution of the continuous-time workload process $V(t)$.

Indeed, (2) states that the expected upward jump size $\mathbb{E}[FX]$ of $\mathbf{V}$, as long as the process stays above b, is smaller than the expected interjump time λ^{-1} so that $\mathbf{V}$ will almost surely return to b after every upcrossing of b within finite time. Thus taking the times between visits of b as cycles, $\mathbf{V}$ is a regenerative process. If the cycle length has finite mean, the ergodic theorem for those processes yields the existence of a limiting steady-state distribution [4]. Due to the negative drift above b, the expected cycle length is finite if the excess of $\mathbf{V}$ over b when entering (b, ∞) has finite mean. This is trivially true if X is almost surely bounded. But if X is not bounded with probability 1, then conditioning on the position of $\mathbf{V}$ just before a jump above level b it is seen that the expected excess over b is bounded by

$$\sup_{0 \leq s \leq b} \mathbb{E}[X - s \mid X > s] \leq \sup_{0 \leq s \leq b} \frac{\mathbb{E}[XI(X > s)]}{\mathbb{P}(X > s)} \leq \frac{\mathbb{E}[X]}{\mathbb{P}(X > b)} < \infty.$$

By $I(A)$ we denote the indicator function of the event A.

QRA is a new variation of queueing models with workload-dependent admission policies. In the literature several models of this kind have been studied:

(a) In finite dam models all workload is accepted up to the capacity restriction b whereas additional workload just overflows and is lost. A comprehensive account of this system for interarrival and service time distributions with rational Laplace-Stieltjes transforms (LSTs) is given in [10], Ch. III.5. The method is based on Pollaczek's contour integral equation which, in the case of rational LSTs, leads to explicit, albeit very complicated formulas. For example, the main results of [10] concerning busy periods in the cases of $M/G/1$ and $G/M/1$ are eqs. (5.82), p. 536, and (5.105), p. 544, respectively, which give joint transforms of the duration of a busy period and two related quantities in terms of contour integrals. In [17] the steady-state workload distribution of the $M/G/1$ model with finite capacity is shown to be proportional to its infinite-capacity counterpart; in [7] this result is extended to workload-dependent rates. Another approach in which the LST of the busy period in the $M/G/1$ case is expressed directly in terms of certain transforms of the underlying distributions is expounded in [27]. Exploiting a certain duality, this technique also yields analogous results for the $G/M/1$ case.

(b) In customer impatience models an arriving customer leaves without receiving any service if his waiting time is too large. Many authors deal with the steady-state distribution of the workload in this model [11, 9, 23, 15, 18, 17]. More recently, various closed-form results were derived [25, 27, 28, 7, 22].

(c) In limited sojourn time models no customer joins the queue whose sojourn time would be too large. Partial results are known for $M/M/1$, $M/D/1$ [9, 15] and $PH/PH/1$ [23].

Due to its importance, e.g. for call centers, the modeling of restricted accessibility and customer impatience has recently attracted a lot of attention [21, 14, 38]. For the general framework of queues with workload-dependent arrival and service rates we refer to [16, 8, 7] and for restricted queues to [34, 37, 20, 25].

In [36] explicit results on the waiting time of a related multi-server system were obtained. In [6] i.i.d. random thresholds, so-called patience times, limiting the time in line for every customer were considered; structural properties of the actual and virtual waiting times in the $GI/G/1$ case and explicit results for Poisson input and patience times with rational Laplace transforms were derived.

In this paper the following results for $M/G/1$ queues with QRA are proved. In Section 2 we derive the steady-state distribution of the workload process, which is also the steady-state distribution of the waiting time, and provide explicit formulas for the cases of Erlang or hyperexponential service requirements and uniformly distributed or constant fractions. In Section 3 we deal with the case of exponential barriers b (instead of one constant threshold). In Sections 4-6 we determine the distribution function of the length of a busy period in closed form. In Sections 7-9 the cycle maximum of the workload process is considered. First we give a new proof for a well-known expression for its distribution function in the unrestricted case, thereby establishing an interesting connection to the density of the associated compound Poisson process, and then derive an alternative formula. We set out to find the cycle maximum distribution in the case when all customers arriving while the workload is below b get full service, whereas no customers are admitted as long as the workload is above b (this is the case $F_n \equiv 1$). Finally we compute the distribution under general QRA. An alternative (martingale) approach to the busy period distribution is presented in Section 10; we illustrate this method for Erlang($2,\mu$)-distributed service requirements and uniform random fractions. In Section 11 we show how the busy period results can be carried over to the non-Markovian case of phase-type arrivals by exploiting a duality between $M/G/1$ and $G/M/1$ type systems. For simplicity we present this approach only for Erlang($2,\mu$)-distributed interarrival times.

2 Waiting times

In this section we shall study the waiting time distribution in the $M/G/1$ queue with QRA as defined above. Let W be a random variable whose distribution is the limiting distribution of W_n for $n \to \infty$. We assume that the stability condition $\lambda\mathbb{E}[FX] < 1$ holds, so that this limiting distribution exists, which because of PASTA is also equal to the steady-state workload distribution in our model.

Starting from the fundamental recursion $W_{n+1} = \max(0, W_n + S_n - A_n)$, $n = 1, 2, \dots,$ we can write:

$$\begin{aligned}
\mathbb{E}[\mathrm{e}^{-\theta W_{n+1}}] &= \mathbb{E}[\mathrm{e}^{-\theta(W_n+S_n-A_n)} I(W_n + S_n > A_n)] + \mathbb{E}[I(W_n + S_n \le A_n)] \\
&= \mathbb{E}[\mathrm{e}^{-\theta(W_n+S_n-A_n)}] - \mathbb{E}[\mathrm{e}^{-\theta(W_n+S_n-A_n)} I(W_n + S_n \le A_n)] \\
&\qquad + \mathbb{E}[I(W_n + S_n \le A_n)] \\
&= \frac{\lambda}{\lambda-\theta}\mathbb{E}[\mathrm{e}^{-\theta(W_n+S_n)}] - \frac{\lambda}{\lambda-\theta}\mathbb{P}(W_n + S_n \le A_n) + \mathbb{P}(W_n + S_n \le A_n) \\
&= \frac{\lambda}{\lambda-\theta}\mathbb{E}[\mathrm{e}^{-\theta(W_n+S_n)}] - \frac{\theta}{\lambda-\theta}\mathbb{P}(W_{n+1} = 0), \quad \mathrm{Re}\,\theta \ge 0. \qquad (3)
\end{aligned}$$

For the limiting distribution of the sojourn time $W_n + S_n$ we obtain from (1) the relation

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}[\mathrm{e}^{-\theta(W_n+S_n)}] &= \int_{x=0}^{b} \mathrm{e}^{-\theta x} \mathrm{d}\mathbb{P}(W < x) \int_{y=0}^{b-x} \mathrm{e}^{-\theta y} \mathrm{d}\mathbb{P}(X < y) \\
&\quad + \int_{x=0}^{b} \mathrm{e}^{-\theta x} \mathrm{d}\mathbb{P}(W < x) \int_{y=b-x}^{\infty} \mathrm{e}^{-\theta(b-x)} \mathbb{E}[\mathrm{e}^{-\theta(y+x-b)F}] \mathrm{d}\mathbb{P}(X < y) \\
&\quad + \int_{x=b}^{\infty} \mathrm{e}^{-\theta x} \mathrm{d}\mathbb{P}(W < x) \mathbb{E}[\mathrm{e}^{-\theta FX}] \\
&=: I + II + III. \qquad (4)
\end{aligned}$$

Introducing the restricted transform

$$a(\theta) := \int_{x=0}^{b} \mathrm{e}^{-\theta x} \mathrm{d}\mathbb{P}(W < x), \qquad (5)$$

we can express the last integral in (4) in the form

$$III = (\mathbb{E}[\mathrm{e}^{-\theta W}] - a(\theta))\mathbb{E}[\mathrm{e}^{-\theta FX}]. \qquad (6)$$

Combining (3), (4), and (6) yields

$$\begin{aligned}
\mathbb{E}[\mathrm{e}^{-\theta W}] &= [1 - \frac{\lambda}{\lambda-\theta}\mathbb{E}[\mathrm{e}^{-\theta FX}]]^{-1} \\
&\qquad \times \Big(- \frac{\theta}{\lambda-\theta}\mathbb{P}(W = 0) + \frac{\lambda}{\lambda-\theta}\Big[I + II - a(\theta)\mathbb{E}[\mathrm{e}^{-\theta FX}]\Big]\Big). \qquad (7)
\end{aligned}$$

The following remark plays an important role in our analysis; in particular, it leads to the determination of $a(\theta)$.

Remark 1. For $0 < x < b$ the waiting time distribution in the QRA model is proportional to the waiting time distribution in the $M/G/1$ queue with service requirements X_n and full accessibility:

$$\mathbb{P}(W < x) = \frac{\mathbb{P}(W = 0)}{1 - \lambda\mathbb{E}X}\mathbb{P}(W < x)_{M/G/1}, \quad 0 < x < b. \qquad (8)$$

Indeed, as long as W_n stays below b, both models behave exactly the same; and when the waiting time in our model returns below level b, the memoryless property of the exponentially distributed interarrival times implies that the $\{W_n\}$ process, restricted to $[0, b]$, behaves the same as the corresponding process for the ordinary unrestricted $M/G/1$ queue when it returns below b. So in both models a busy period probabilistically has as many customers with waiting time less than x, for $0 < x < b$. Using the theory of regenerative processes, one can write

$$\mathbb{P}(W < x) \;=\; \frac{1}{\mathbb{E}N}\mathbb{E}\Big[\sum_{j=1}^{N} I(W_j < x)\Big],$$

with N the number of customers served in a busy period. Hence, for $0 < x < b$, the proportionality factor between $\mathbb{P}(W < x)$ and $\mathbb{P}(W < x)_{M/G/1}$ is the ratio of the mean numbers of customers in a busy period of both systems, which equals the ratio of $\mathbb{P}(W = 0) =: \Pi_0$ in the QRA model and $\mathbb{P}(W = 0) = 1 - \lambda\mathbb{E}X$ in the ordinary $M/G/1$ queue. ■

It follows from Remark 1 that $\mathbb{P}(W < x)$, for $0 < x < b$, and thus also $a(\theta)$ are known up to the unknown constant Π_0. Note that the integrals I and II are expressed by means of the distributions of X and F and of $\mathbb{P}(W < x)$, $0 < x < b$; therefore, $\mathbb{E}[\mathrm{e}^{-\theta W}]$ is determined by (7) once we know Π_0. This constant can be computed from the familiar normalization argument: $\mathbb{E}[\mathrm{e}^{-\theta W}] = 1$ for $\theta = 0$. However, the following should be realized. While the LST of W is well-known for the standard $M/G/1$ queue, the distribution function of W in this case can in general only be given in terms of an inconvenient convolution series:

$$\mathbb{P}(W < x)_{M/G/1} = (1 - \lambda\mathbb{E}X)\sum_{k=0}^{\infty} H^{(k)}(x),$$

where $H^{(k)}$ is the $k$fold convolution of the function $H(x) = \lambda\int_0^x \mathbb{P}(X > u)\,\mathrm{d}u$ with itself. It *is* known explicitly for the class of service requirement distributions with a rational LST (cf. Section II.5.10 of [10]). To make our results more explicit, we restrict ourselves to that class in the remainder of this section. This is not a major restriction as the class of such distributions is dense in the class of all distributions of non-negative random variables. It is well-known (see Formula (II.5.190) of [10]) that the LST of the steady-state waiting time of the ordinary $M/G/1$ queue with a service time LST $\beta(\theta) = \beta_1(\theta)/\beta_2(\theta)$, where $\beta_2(\theta)$ is a polynomial in θ of degree k and $\beta_1(\theta)$ is a polynomial of degree at most $k-1$, is given by

$$\mathbb{E}[\mathrm{e}^{-\theta W}]_{M/G/1} \;=\; \frac{\beta_2(\theta)}{\beta_2(0)}\prod_{i=1}^{k}\frac{\xi_i}{\xi_i - \theta}, \quad \mathrm{Re}\,\theta \geq 0. \tag{9}$$

Here $\xi_1, \ldots, \xi_k$ are the k zeros of $\beta_2(\theta) - \beta_1(\theta)\frac{\lambda}{\lambda - \theta}$ in the complex left half-plane, each occurring with its multiplicity. It follows that the distribution of W is a mixture of convolutions of exponential distributions in which some of the weights may be negative. For simplicity let us assume for the moment that all the ξ_i are distinct. Then by (9),

$$\mathbb{P}(W > y)_{M/G/1} = \sum_{i=1}^{k} C_i \mathrm{e}^{\xi_i y}, \quad y \geq 0. \tag{10}$$

Hence in the QRA model we have, for $0 < x < b$,

$$\mathbb{P}(W < x) = \frac{\Pi_0}{1 - \lambda\mathbb{E}X}\Big(1 - \sum_{i=1}^{k} C_i \mathrm{e}^{\xi_i x}\Big) = \Pi_0\frac{1 - \sum_{i=1}^{k} C_i \mathrm{e}^{\xi_i x}}{1 - \sum_{i=1}^{k} C_i} \tag{11}$$

and

$$a(\theta) = \int_{x=0}^{b} \mathrm{e}^{-\theta x} \mathrm{d}\mathbb{P}(W < x) = \frac{\Pi_0}{1-\sum_{i=1}^{k} C_i} \frac{\sum_{i=1}^{k} C_i(-\xi_i)(1-\mathrm{e}^{-(\theta-\xi_i)b})}{\theta-\xi_i}. \qquad (12)$$

To demonstrate the approach towards evaluating the terms I and II on the righthand side of (4), we shall restrict ourselves in the calculations to two classes of distributions with rational LSTs:

(i) Erlang distributions;

(ii) hyperexponential distributions.

By taking appropriate weighted combinations, one can also handle the more general case of rational LST.

In addition, we shall consider two choices for the distribution of F:

(a) F is uniformly distributed on $(0,1)$;

(b) F is constant.

This yields a collection of four cases, which will be discussed consecutively.

Case 1: *$F = U$, with U uniformly distributed on $(0,1)$, and Erlang(k,μ)-distributed service requirements.*
Observe that

$$\mathbb{E}[\mathrm{e}^{-\theta U X}] = \int_{0}^{\infty} \mathbb{E}[\mathrm{e}^{-\theta U y}] \mathrm{d}\mathbb{P}(X < y) = \int_{0}^{\infty} \frac{1-\mathrm{e}^{-\theta y}}{\theta y} \mathrm{d}\mathbb{P}(X < y). \qquad (13)$$

In this case, cf. (4),

$$\begin{aligned} I + II &= \int_{x=0}^{b} \mathrm{e}^{-\theta x} \mathrm{d}\mathbb{P}(W < x) \int_{y=0}^{b-x} \mathrm{e}^{-\theta y} \mu \frac{(\mu y)^{k-1}}{(k-1)!} \mathrm{e}^{-\mu y} \mathrm{d}y \\ &\quad + \int_{x=0}^{b} \mathrm{e}^{-\theta x} \mathrm{d}\mathbb{P}(W < x) \int_{y=b-x}^{\infty} \mathrm{e}^{-\theta(b-x)} \frac{1-\mathrm{e}^{-\theta(y+x-b)}}{\theta(y+x-b)} \mu \frac{(\mu y)^{k-1}}{(k-1)!} \mathrm{e}^{-\mu y} \mathrm{d}y. \qquad (14) \end{aligned}$$

First consider term I:

$$\begin{aligned} I &= \int_{x=0}^{b} \mathrm{e}^{-\theta x} \mathrm{d}\mathbb{P}(W < x) (\frac{\mu}{\mu+\theta})^k \Big[1 - \sum_{j=0}^{k-1} \mathrm{e}^{-(\mu+\theta)(b-x)} \frac{((\mu+\theta)(b-x))^j}{j!}\Big] \\ &= (\frac{\mu}{\mu+\theta})^k a(\theta) - (\frac{\mu}{\mu+\theta})^k \sum_{j=0}^{k-1} \frac{(\mu+\theta)^j}{j!} \mathrm{e}^{-(\mu+\theta)b} \int_{x=0}^{b} \mathrm{e}^{\mu x} (b-x)^j \mathrm{d}\mathbb{P}(W < x). \qquad (15) \end{aligned}$$

Notice that the last integral can be easily evaluated, since the distribution of W is given by the simple expression (11). Since $a(\theta)$ is easily computed by specializing (12) to the Erlang(k, μ)-case, we see that (15) yields I in closed form as a linear combination of simple rational and exponential functions of θ. We refrain from giving the details, and turn to term II:

$$
\begin{aligned}
II &= \int_{x=0}^{b} \mathrm{e}^{-\theta x}\mathrm{d}\mathbb{P}(W<x)\int_{z=0}^{\infty}\mathrm{e}^{-\theta(b-x)}\frac{1-\mathrm{e}^{-\theta z}}{\theta z}\mu\frac{\mu^{k-1}}{(k-1)!}\mathrm{e}^{-\mu(z+b-x)}(z+b-x)^{k-1}\mathrm{d}z \\
&= \mu\frac{\mu^{k-1}}{(k-1)!}\mathrm{e}^{-(\mu+\theta)b}\int_{x=0}^{b}\mathrm{e}^{\mu x}\mathrm{d}\mathbb{P}(W<x) \\
&\quad \times\int_{z=0}^{\infty}\frac{1-\mathrm{e}^{-\theta z}}{\theta z}\mathrm{e}^{-\mu z}((b-x)^{k-1}+\sum_{m=1}^{k-1}\binom{k-1}{m}z^m(b-x)^{k-1-m})\mathrm{d}z \\
&= \mu\frac{\mu^{k-1}}{(k-1)!}\mathrm{e}^{-(\mu+\theta)b}\int_{x=0}^{b}\mathrm{e}^{\mu x}\mathrm{d}\mathbb{P}(W<x) \qquad (16)\\
&\quad \times\Big((b-x)^{k-1}\int_{z=0}^{\infty}\frac{1-\mathrm{e}^{-\theta z}}{\theta z}\mathrm{e}^{-\mu z}\mathrm{d}z \qquad (17)\\
&\qquad +\sum_{m=1}^{k-1}\binom{k-1}{m}(b-x)^{k-1-m}\int_{z=0}^{\infty}\frac{1-\mathrm{e}^{-\theta z}}{\theta}\mathrm{e}^{-\mu z}z^{m-1}\mathrm{d}z\Big).
\end{aligned}
$$

Hence we can write

$$
II=\mu\frac{\mu^{k-1}}{(k-1)!}\mathrm{e}^{-(\mu+\theta)b}\sum_{m=0}^{k-1}f_m(\theta)\int_{x=0}^{b}(b-x)^{k-1-m}\mathrm{e}^{\mu x}\mathrm{d}\mathbb{P}(W<x), \qquad (18)
$$

where $f_0(\theta),\dots,f_{k-1}(\theta)$ are known functions. The most complicated one is $f_0(\theta)$, known as the Frullani integral:

$$
\begin{aligned}
f_0(\theta) &= \int_{z=0}^{\infty}\frac{1-\mathrm{e}^{-\theta z}}{\theta z}\mathrm{e}^{-\mu z}\mathrm{d}z=\frac{1}{\theta}\int_{z=0}^{\infty}(1-\mathrm{e}^{-\theta z})\mathrm{d}z\int_{y=\mu}^{\infty}\mathrm{e}^{-yz}\mathrm{d}y \\
&= \frac{1}{\theta}\int_{y=\mu}^{\infty}(\frac{1}{y}-\frac{1}{\theta+y})\mathrm{d}y=\frac{1}{\theta}\ln(1+\frac{\theta}{\mu}). \qquad (19)
\end{aligned}
$$

For $m=1,\dots,k-1$, $f_m(\theta)$ is given by

$$
f_m(\theta)=\binom{k-1}{m}\frac{1}{\theta}\int_{z=0}^{\infty}(1-\mathrm{e}^{-\theta z})\mathrm{e}^{-\mu z}z^{m-1}\mathrm{d}z=\binom{k-1}{m}\frac{1}{\theta}\Big(\frac{(m-1)!}{\mu^m}-\frac{(m-1)!}{(\mu+\theta)^m}\Big). \qquad (20)
$$

Combining (7), (15) and (18), we arrive at

$$
\begin{aligned}
\mathbb{E}[\mathrm{e}^{-\theta W}] &= \Big[1-\frac{\lambda}{\lambda-\theta}\mathbb{E}[\mathrm{e}^{-\theta UX}]\Big]^{-1} \\
&\quad \times\Big\{-\frac{\theta}{\lambda-\theta}\mathbb{P}(W=0)+\frac{\lambda}{\lambda-\theta}\Big[-a(\theta)\mathbb{E}[\mathrm{e}^{-\theta UX}]+a(\theta)(\frac{\mu}{\mu+\theta})^k \\
&\quad -(\frac{\mu}{\mu+\theta})^k\sum_{j=0}^{k-1}\frac{(\mu+\theta)^j}{j!}\mathrm{e}^{-(\mu+\theta)b}\int_{x=0}^{b}\mathrm{e}^{\mu x}(b-x)^j\mathrm{d}\mathbb{P}(W<x) \\
&\quad +\mu\frac{\mu^{k-1}}{(k-1)!}\mathrm{e}^{-(\mu+\theta)b}\sum_{m=0}^{k-1}f_m(\theta)\int_{x=0}^{b}(b-x)^{k-1-m}\mathrm{e}^{\mu x}\mathrm{d}\mathbb{P}(W<x)\Big]\Big\}. \qquad (21)
\end{aligned}
$$

Notice that $\mathbb{E}[\mathrm{e}^{-\theta UX}]$ is in the Erlang(k,μ) case given by $\sum_{j=1}^{k-1}(\mu/(\mu+\theta))^j/(k-1)$. The only remaining unknown on the righthand side of (26) is $\mathbb{P}(W=0)$; it also appears in

$\mathbb{P}(W < x)$, $0 < x < b$ (recall Remark 1). This probability follows via a normalization argument: $\mathbb{E}[\mathrm{e}^{-\theta W}] = 1$ for $\theta = 0$. If we denote by $W + S$ a steady-state random variable for the sojourn time, then $\mathbb{P}(W = 0) = \mathbb{E}[\mathrm{e}^{-\lambda(W+S)}]$, as the righthand term is the steady-state probability that no arrival takes place during the sojourn time of a customer.

Case 2: $F = U$; *hyperexponentially distributed service requirements.*
Let $\mathbb{P}(X > y) = \sum_{j=1}^{k} b_j \mathrm{e}^{-\mu_j y}$ with positive constants $b_1, \dots, b_k$ adding up to 1. In this case, term I becomes:

$$\begin{aligned} I &= \int_{x=0}^{b} \mathrm{e}^{-\theta x} \mathrm{d}\mathbb{P}(W < x) \sum_{j=1}^{k} \frac{b_j \mu_j}{\mu_j + \theta} (1 - \mathrm{e}^{-(\mu_j+\theta)(b-x)}) \\ &= \sum_{j=1}^{k} \frac{b_j \mu_j}{\mu_j + \theta} [a(\theta) - \mathrm{e}^{-(\mu_j+\theta)b} a(-\mu_j)]. \end{aligned} \tag{22}$$

A straightforward calculation yields the following expression for term II (one could also take $k = 1$ in (18), multiply by b_j and take the sum over j):

$$II = \sum_{j=1}^{k} a(-\mu_j) b_j \mu_j \mathrm{e}^{-(\mu_j+\theta)b} \frac{1}{\theta} \ln(1 + \frac{\theta}{\mu_j}). \tag{23}$$

Combining (7), (22) and (23) yields

$$\begin{aligned} \mathbb{E}[\mathrm{e}^{-\theta W}] &= \Big[1 - \frac{\lambda}{\lambda - \theta} \mathbb{E}[\mathrm{e}^{-\theta U X}]\Big]^{-1} \\ &\quad \times \Big\{ - \frac{\theta}{\lambda - \theta} \mathbb{P}(W = 0) + \frac{\lambda}{\lambda - \theta} \Big[- a(\theta) \mathbb{E}[\mathrm{e}^{-\theta U X}] + a(\theta) \sum_{j=1}^{k} \frac{b_j \mu_j}{\mu_j + \theta} \\ &\quad + \sum_{j=1}^{k} b_j \mu_j a(-\mu_j) \mathrm{e}^{-(\mu_j+\theta)b} [\frac{1}{\theta} \ln(1 + \frac{\theta}{\mu_j}) - \frac{1}{\mu_j + \theta}]\Big]\Big\}. \end{aligned} \tag{24}$$

Notice that $\mathbb{E}[\mathrm{e}^{-\theta U X}]$ is in the hyperexponential case given by $\sum_{j=1}^{k} b_j \mu_j \theta^{-1} \ln(1 + (\theta/\mu_j))$, and that $\mathbb{P}(W = 0)$ is again obtained via normalization.

Case 3: $F \equiv d$, *where d is a constant,* $0 < d < 1$*; Erlang*(k, μ)*-distributed service requirements.*
Term I is the same as in (15). Term II becomes

$$\begin{aligned} II &= \int_{x=0}^{b} \mathrm{e}^{-\theta x} \mathrm{d}\mathbb{P}(W < x) \int_{y=b-x}^{\infty} \mathrm{e}^{-\theta(b-x)} \mathrm{e}^{-\theta d(y+x-b)} \mathrm{d}\mathbb{P}(X < y) \\ &= \mathrm{e}^{-\theta(1-d)b} \int_{x=0}^{b} \mathrm{e}^{-\theta d x} \mathrm{d}\mathbb{P}(W < x) \int_{y=b-x}^{\infty} \mathrm{e}^{-\theta d y} \mu \frac{(\mu y)^{k-1}}{(k-1)!} \mathrm{e}^{-\mu y} \mathrm{d}y \\ &= \mathrm{e}^{-\theta(1-d)b} \int_{x=0}^{b} \mathrm{e}^{-\theta d x} \mathrm{d}\mathbb{P}(W < x) \Big[(\frac{\mu}{\mu + \theta d})^k \sum_{j=0}^{k-1} \mathrm{e}^{-(\mu+\theta d)(b-x)} \frac{((\mu + \theta d)(b - x))^j}{j!} \Big] \\ &= (\frac{\mu}{\mu + \theta d})^k \sum_{j=0}^{k-1} \frac{(\mu + \theta d)^j}{j!} \mathrm{e}^{-(\mu+\theta)b} \int_{x=0}^{b} \mathrm{e}^{\mu x} (b - x)^j \mathrm{d}\mathbb{P}(W < x). \end{aligned} \tag{25}$$

Notice that for $d = 1$ this last expression coincides with the last part of (15), as could be expected.

Combining (7), (15) and (25) yields

$$\begin{aligned}
\mathbb{E}[\mathrm{e}^{-\theta W}] &= \Big[1 - \frac{\lambda}{\lambda-\theta}(\frac{\mu}{\mu+\theta d})^k\Big]^{-1} \\
&\times \Big\{ - \frac{\theta}{\lambda-\theta}\mathbb{P}(W=0) + \frac{\lambda}{\lambda-\theta}\Big[- a(\theta)(\frac{\mu}{\mu+\theta d})^k + a(\theta)(\frac{\mu}{\mu+\theta})^k \\
&- (\frac{\mu}{\mu+\theta})^k \sum_{j=0}^{k-1} \frac{(\mu+\theta)^j}{j!}\mathrm{e}^{-(\mu+\theta)b} \int_{x=0}^{b} \mathrm{e}^{\mu x}(b-x)^j \mathrm{d}\mathbb{P}(W<x) \\
&+ (\frac{\mu}{\mu+\theta d})^k \sum_{j=0}^{k-1} \frac{(\mu+\theta d)^j}{j!}\mathrm{e}^{-(\mu+\theta)b} \int_{x=0}^{b} \mathrm{e}^{\mu x}(b-x)^j \mathrm{d}\mathbb{P}(W<x)\Big]\Big\}. \qquad (26)
\end{aligned}$$

Case 4: $F \equiv d$; *hyperexponentially distributed service requirements.*
Again let $\mathbb{P}(X > y) = \sum_{j=1}^{k} b_j \mathrm{e}^{-\mu_j y}$. In this case, term I is already given by (22). For term II we obtain

$$\begin{aligned}
II &= \int_{x=0}^{b} \mathrm{e}^{-\theta x}\mathrm{d}\mathbb{P}(W<x) \int_{y=b-x}^{\infty} \mathrm{e}^{-\theta(b-x)}\mathrm{e}^{-\theta d(y+x-b)}\mathrm{d}\mathbb{P}(X<y) \\
&= \mathrm{e}^{-\theta(1-d)b} \int_{x=0}^{b} \mathrm{e}^{-\theta d x}\mathrm{d}\mathbb{P}(W<x) \int_{y=b-x}^{\infty} \mathrm{e}^{-\theta d y} \sum_{j=1}^{k} b_j\mu_j \mathrm{e}^{-\mu_j y}\mathrm{d}y \\
&= \mathrm{e}^{-(\mu_j+\theta)b} \sum_{j=1}^{k} \frac{b_j\mu_j}{\mu_j+\theta d} a(-\mu_j). \qquad (27)
\end{aligned}$$

Combining (7), (22) and (27) yields:

$$\begin{aligned}
\mathbb{E}[\mathrm{e}^{-\theta W}] &= \Big[1 - \frac{\lambda}{\lambda-\theta}\sum_{j=1}^{k}\frac{b_j\mu_j}{\mu_j+\theta d}\Big]^{-1} \qquad (28) \\
&\times\Big\{ - \frac{\theta}{\lambda-\theta}\mathbb{P}(W=0) + \frac{\lambda}{\lambda-\theta}\sum_{j=1}^{k}(\frac{b_j\mu_j}{\mu_j+\theta} - \frac{b_j\mu_j}{\mu_j+\theta d})(a(\theta) - \mathrm{e}^{-(\mu_j+\theta)b}a(-\mu_j))\Big\}.
\end{aligned}$$

We end this section with the following remarks.

Remark 2. Now that the distribution of W has been determined, we can also find the steady-state distribution of the *sojourn* time $R := W + S$. It should be observed that W and S are dependent, cf. (1), and that $\mathbb{E}[\mathrm{e}^{-\theta R}] = I + II + III$, cf. (4). Terms I and II are evaluated in detail in the four cases discussed above, and III is given in (6). Finally, the steady-state distribution of the number of customers right after a departure, Z, is given by the relation

$$\mathbb{E}[\mathrm{e}^{-\lambda(1-r)R}] = \mathbb{E}[r^Z].$$

As in an ordinary $M/G/1$ queue, Z has the same distribution as the number of customers seen by an arrival, and also (by PASTA) as the number of customers in steady state. ■

Remark 3. In principle, the method developed in this section also allows us to handle the case in which, above level b, there are additional levels b_2, b_3 etc. above which different fractions of the surplus service requirements are admitted. Such models would behave in the same way (up to a proportionality constant) *below* level b_2, so one can restrict oneself to analyzing W above b_2, and so on. ■

3 A model variant: the case of exponential patience

In the previous section, a fixed barrier b determined whether a service requirement was fully or only partially fulfilled. In the present section we consider the case of exponential patience, viz., the barrier B_n for the nth customer is exponentially distributed with mean, say, $1/\zeta$. The actual service time S_n of the nth customer depends on the service requirement X_n and the waiting time W_n in the same way as in (1), but b in that formula now becomes an exp(ζ)-distributed random variable (B_n for the nth customer). To study the steady-state waiting time distribution in this case, we shall integrate the expressions for I, II and III in the previous section with respect to the density of B_n. From (6) we obtain III for the new model:

$$\begin{aligned} III &= \mathbb{E}[\mathrm{e}^{-\theta FX}]\Big(\mathbb{E}[\mathrm{e}^{-\theta W}] - \int_{b=0}^{\infty} \zeta \mathrm{e}^{-\zeta b}\mathrm{d}b \int_{x=0}^{b} \mathrm{e}^{-\theta x}\mathrm{d}\mathbb{P}(W < x)\Big) \\ &= \mathbb{E}[\mathrm{e}^{-\theta FX}]\{\mathbb{E}[\mathrm{e}^{-\theta W}] - \mathbb{E}[\mathrm{e}^{-(\theta+\zeta)W}]\}. \end{aligned} \tag{29}$$

The above formula is easily interpreted: the term between curly brackets equals $\mathbb{E}[\mathrm{e}^{-\theta W} I(W > B)]$, as could have been expected from the last part of (1). (Here B is a new generic exp(ζ)-distributed random variable which is independent of W and X.)

The term I becomes, after a lengthy calculation or after the simple observation that $\mathbb{E}[\mathrm{e}^{-\theta(W+X)} I(W + X \leq B)] = \mathbb{E}[\mathrm{e}^{-(\theta+\zeta)(W+X)}]$ (cf. the first part of (1)):

$$I = \mathbb{E}[\mathrm{e}^{-(\theta+\zeta)(W+X)}]. \tag{30}$$

Finally we determine II, performing some interchanges of integrations:

$$\begin{aligned} II &= \int_{b=0}^{\infty} \zeta \mathrm{e}^{-\zeta b}\mathrm{d}b \int_{x=0}^{b} \mathrm{e}^{-\theta x}\mathrm{d}\mathbb{P}(W < x) \int_{y=b-x}^{\infty} \mathrm{e}^{-\theta(b-x)}\mathbb{E}[\mathrm{e}^{-\theta(y+x-b)F}]\mathrm{d}\mathbb{P}(X < y) \\ &= \mathbb{E}[\mathrm{e}^{-(\theta+\zeta)W}] \int_{v=0}^{\infty} \zeta \mathrm{e}^{-(\theta+\zeta)v} \int_{y=v}^{\infty} \mathbb{E}[\mathrm{e}^{-\theta(y-v)F}]\mathrm{d}\mathbb{P}(X < y)\mathrm{d}v. \end{aligned} \tag{31}$$

Combining (3), (4), (29), (30) and (31) yields:

$$\begin{aligned} \mathbb{E}[\mathrm{e}^{-\theta W}] &= \Big[1 - \frac{\lambda}{\lambda-\theta}\mathbb{E}[\mathrm{e}^{-\theta FX}]\Big]^{-1} \\ &\quad \times\Big\{ - \frac{\theta}{\lambda-\theta}\mathbb{P}(W=0) + \frac{\lambda}{\lambda-\theta}\Big[- \mathbb{E}[\mathrm{e}^{-\theta FX}]\mathbb{E}[\mathrm{e}^{-(\theta+\zeta)W}] \\ &\quad +\mathbb{E}[\mathrm{e}^{-(\theta+\zeta)X}]\mathbb{E}[\mathrm{e}^{-(\theta+\zeta)W}] \\ &\quad +\mathbb{E}[\mathrm{e}^{-(\theta+\zeta)W}] \int_{v=0}^{\infty} \zeta \mathrm{e}^{-(\theta+\zeta)v} \int_{y=v}^{\infty} \mathbb{E}[\mathrm{e}^{-\theta(y-v)F}]\mathrm{d}\mathbb{P}(X < y)\mathrm{d}v\Big]\Big\}. \end{aligned} \tag{32}$$

For the four cases discussed above, it is easy to evaluate the double integral on the righthand side of (32). We will not spell out the details. More importantly, it should be observed that (32) has the following structure:

$$\mathbb{E}[\mathrm{e}^{-\theta W}] = f_1(\theta)\mathbb{P}(W=0) + f_2(\theta)\mathbb{E}[\mathrm{e}^{-(\theta+\zeta)W}], \tag{33}$$

$f_1(\theta)$ and $f_2(\theta)$ being functions which are expressed in known quantities:

$$f_1(\theta) := \frac{\theta}{\theta - \lambda + \lambda\mathbb{E}[\mathrm{e}^{-\theta F X}]}, \tag{34}$$

$$\begin{aligned} f_2(\theta) \;\; := \;\; & \frac{\theta-\lambda}{\theta - \lambda + \lambda\mathbb{E}[\mathrm{e}^{-\theta F X}]}\Big[-\frac{\lambda}{\lambda-\theta}\mathbb{E}[\mathrm{e}^{-\theta F X}] + \mathbb{E}[\mathrm{e}^{-(\theta+\zeta)X}] \\ & + \int_{v=0}^{\infty} \zeta \mathrm{e}^{-(\theta+\zeta)v} \int_{y=v}^{\infty} \mathbb{E}[\mathrm{e}^{-\theta(y-v)F}]\mathrm{d}\mathbb{P}(X<y)\mathrm{d}v\Big]. \end{aligned} \tag{35}$$

Formally, iteration leads to

$$\mathbb{E}[\mathrm{e}^{-\theta W}] = \mathbb{P}(W=0)\sum_{j=0}^{\infty} f_1(\theta+j\zeta)\prod_{k=0}^{j-1} f_2(\theta+k\zeta). \tag{36}$$

Normalization finally determines $\mathbb{P}(W=0)$, leading to the following expression for the LST of the steady-state waiting time (and workload) in the model with exponential “patience”:

$$\mathbb{E}[\mathrm{e}^{-\theta W}] = \frac{\sum_{j=0}^{\infty} f_1(\theta+j\zeta)\prod_{k=0}^{j-1} f_2(\theta+k\zeta)}{\sum_{j=0}^{\infty} f_1(j\zeta)\prod_{k=0}^{j-1} f_2(k\zeta)}. \tag{37}$$

To prove (36) and thus (37), consider the Nth iteration of (33):

$$\begin{aligned} \mathbb{E}[\mathrm{e}^{-\theta W}] = \mathbb{P}(W=0)\sum_{j=0}^{N} f_1(\theta+j\zeta)\prod_{k=0}^{j-1} f_2(\theta+k\zeta) \\ + \Big(\prod_{k=0}^{N} f_2(\theta+k\zeta)\Big)\mathbb{E}[\mathrm{e}^{-(\theta+(N+1)\zeta)W}]. \end{aligned} \tag{38}$$

Next note that $f_1(\theta) > 0$ and $f_2(\theta) > 0$ for all $\theta > \lambda$ and that

$$\lim_{j\to\infty} f_1(\theta+j\zeta) = 1. \tag{39}$$

It thus follows from (38) that

$$\mathbb{E}[\mathrm{e}^{-\theta W}] \geq \mathbb{P}(W=0)\sum_{j=0}^{N} f_1(\theta+j\zeta)\prod_{k=0}^{j-1} f_2(\theta+k\zeta) \tag{40}$$

for all $N \in \mathbb{N}$ and all $\theta > \lambda$. Letting $N \to \infty$ in (40) it is seen that for all $\theta > \lambda$

$$0 < \sum_{j=0}^{\infty} f_1(\theta + j\zeta) \prod_{k=0}^{j-1} f_2(\theta + k\zeta) < \infty \tag{41}$$

The Nth summand in this series is positive and converges to zero as $N \to \infty$ so that, by (39),

$$\lim_{j\to\infty} \prod_{k=0}^{j-1} f_2(\theta + k\zeta) = 0. \tag{42}$$

Eq. (42) has been proved for $\theta > \lambda$ but it is obvious that it then also holds for $\theta < \lambda$ as long as $(\lambda - \theta)/\zeta$ is not an integer. It follows that the remainder term $\big(\prod_{k=0}^{N} f_2(\theta + k\zeta)\big)\mathbb{E}[\mathrm{e}^{-(\theta+(N+1)\zeta)W}]$ in (38) tends to zero as $N \to \infty$. This completes the proof of (36).

4 The busy period of the QRA queue

In the following section we derive the distribution of the duration of a busy period of the QRA $M/G/1$ queue. We need the compound Posson process $\{Y(t) \mid t \geq 0\}$ whose jump times are the customer arrival times and whose jump sizes are the service requirements $X_1, X_2, \ldots$ Formally, let

$$Y(t) = \sum_{n=1}^{N(t)} X_n, \quad t \geq 0 \tag{43}$$

where $\{N(t) \mid t \geq 0\}$ is the ordinary Poisson counting process generated by the arrival times, and $\sum_{n=1}^{0} = 0$. We call this process $CPP(\lambda, B)$ where $B(x)$ is the common distribution function of the X_n. We assume that $B(x)$ has a density $f(x)$.

Let $0 < \beta_1, \beta_2 < \infty$ and define the two stopping times

$$T_L(\beta_1) = \inf\{t > 0 : Y(t) = -\beta_1 + t\} \tag{44}$$

and

$$T_U(\beta_2) = \inf\{t > 0 : Y(t) \geq \beta_2 + t\}. \tag{45}$$

If there is no restriction on service capacity, then the length of a busy period, starting with $V(0) = X_0$, is $T_L(X_0)$. In the case of QRA we need to introduce additional stopping times for the various phases during a busy period. In the sequel we will express the LST of the busy period under QRA in terms of transforms for these phases. In Section 5 we derive the various components of the LST formula for the case of an $M/M/1$ queue, and in Section 6 we do the same in the much more complicated general $M/G/1$ case.

Assume that at the beginning of the busy period (which we set at $t = 0$) a customer arrives with a demand for service X_0. We distinguish between two cases:

Case I: $X_0 \leq b$;

Case II: $X_0 > b$.

In Case I, the $V(t)$ process starts at X_0 and moves below the boundary b until for the first time it either hits zero (ending the busy period) or jumps above b. This stopping time is

$$T(X_0, b - X_0) = \min\{T_L(X_0), T_U(b - X_0)\}. \tag{46}$$

The time interval $[0, T(X_0, b - X_0))$ is called the *initial phase.* If $T(X_0, b-X_0) = T_U(b-X_0)$ the initial phase ends with $V(t)$ jumping above b. We denote the overshoot by R_1:

$$R_1 = Y(T_U(b - X_0)) - (b - X_0 + T_U(b - X_0)). \tag{47}$$

In the subsequent over-b-phase, the $V(t)$ process continues from $\beta_1^* = R_1 F_1$. In this phase $V(t)$ behaves like a CPP(λ, B^*) where $B^*(x) = \mathbb{P}(FX \le x)$. The density of B^* is

$$f^*(x) = \int\limits_0^1 \frac{1}{u} f\left(\frac{x}{u}\right) \mathrm{d}\mathbb{P}(F < u). \tag{48}$$

Let $\{Y^*(t), t \ge 0\}$ denote the CPP

$$Y^*(t) = \sum_{n=1}^{N^*(t)} X_n^*, \tag{49}$$

where the density of the i.i.d. X_n^*, $n \ge 1$, is f^*, the ordinary Poisson process $N^*(t)$ is independent of the X_n^*, and the new CPP is independent of $CPP(\lambda, B)$. The length of the over-b-phase is

$$T_L^*(\beta_1^*) = \inf\{t > 0 : Y^*(t) = -\beta_1^* + t\}. \tag{50}$$

Clearly, $T_U(b - X_1) + T_L^*(\beta_1^*)$ can be viewed as the first time instant at which $V(t)$ drops below b. This instant is the end of the *initial cycle.*

In Case II the $V(t)$ process is already above b at time zero. We then define $R_1^* = X_1 - b$, $\beta_1^* = R_1^* F_1$; the initial cycle ends at $T_L^*(\beta_1^*)$.

The end of the initial cycle is a regeneration point for the workload process, which restarts from level b. At this time the $V(t)$ process behaves as in Case I with $X_0 \equiv b$. If after regeneration it hits 0 before jumping above b, the busy period is ended (the corresponding time interval is called a *terminating phase*). If it jumps above b before reaching 0, a new over-b-phase will follow during which the workload returns to b. The time until then is called a *continuing* cycle. The busy period continues until the first terminating phase occurs.

Obviously, the random number N of continuing cycles before a terminating phase has a shifted geometric distribution, i.e.,

$$\mathbb{P}(N = m) = pq^m, \quad m = 0, 1, \ldots \tag{51}$$

where

$$p = \mathbb{P}(T_L(b) < T_U(0)). \tag{52}$$

The above arguments show how the LST $W^*(\cdot \mid b)$ of the length of a busy period is composed of the LSTs of the various phases and cycles:

$$W^*(\theta \mid b) = M_L^{(I)}(\theta \mid b) + \frac{M_U^{(I)}(\theta \mid b)\psi_L^*(\theta \mid b)}{1 - \psi_C^*(\theta \mid b)}, \quad \text{Re}\ \theta \ge 0. \tag{53}$$

The functions on the righthand side of (53) are the following. For the initial phase,

$$M_L^{(I)}(\theta \mid b) = \int_0^b f(x)\mathbb{E}[\mathrm{e}^{-\theta T_L(x)} I(T_L(x) < T_U(b - x))]\ \mathrm{d}x, \tag{54}$$

$$\begin{aligned}M_U^{(I)}(\theta \mid b) = \int_0^b f(x)\mathbb{E}[\mathrm{e}^{-\theta(T_U(b-x)+T_L^*(R_1F_1))}I(T_U(b-x) < T_L(x))]\ \mathrm{d}x \\ + \int_b^\infty f(x)\mathbb{E}[\mathrm{e}^{-\theta T_L^*((x-b)F_1)}]\ \mathrm{d}x.\end{aligned} \tag{55}$$

For a terminating phase starting at b the LST is

$$\psi_L^*(\theta \mid b) = \mathbb{E}[\mathrm{e}^{-\theta T_L(b)}I(T_L(b) < T_U(0))]. \tag{56}$$

Finally, for a continuing cycle the LST is

$$\psi_c^*(\theta \mid b) = \mathbb{E}[\mathrm{e}^{-\theta(T_U(0)+T_L^*(R_1F_1))}I(T_U(0) < T_L(b))]. \tag{57}$$

The LSTs (54)-(57) are determined explicitly in the following sections, first for the $M/M/1$ case and then for $M/G/1$.
Remark. The stopping time $T_L^*(\beta_1^*)$ was defined in (50). For fixed β_1^* its LST can be expressed in terms of the probability density $h^*(\cdot;t)$ of $Y^*(t)$. Clearly, $h^*(y;t)$ is given by

$$h^*(y,t) = \sum_{n=1}^{\infty} \mathrm{e}^{-\lambda t}\frac{(\lambda t)^n}{n!}f^{*(n)}(y), \quad y > 0 \tag{58}$$

where $f^{*(n)}$, $n \geq 1$, is the $n$fold convolution of f^* (which in turn is given in (48)). It is well known (see [5]) that the density of $T_L^*(\beta_1^*)$, say $\psi_L^*(t;\beta_1^*)$, is related to $h^*(y;t)$ by the identity

$$\psi_L^*(t;\beta_1^*) = \frac{\beta_1^*}{t}h^*(t-\beta_1^*;t), \quad t > \beta_1^*. \tag{59}$$

Accordingly,

$$\mathbb{E}[\mathrm{e}^{-\theta T_L^*(\beta_1^*)}] = \mathrm{e}^{-(\lambda+\theta)\beta_1^*} + \beta_1^*\mathrm{e}^{-\theta\beta_1^*}\int_0^\infty \mathrm{e}^{-\theta t}\frac{1}{t+\beta_1^*}h^*(t;\beta_1^*+t)\ \mathrm{d}t. \tag{60}$$

■

5 The $M/M/1$ case

In the $M/M/1$ case with exp(μ)-distributed service requirements there is a significant simplification, since the overshoot R_1 is independent of $T_U(\beta_2)$ and exp(μ)-distributed. Thus, setting $W = R_1F_1$,

$$\begin{aligned}&\mathbb{E}[\mathrm{e}^{-\theta T_U(\beta_2)-\theta T_L^*(W)}I(T_U(\beta_2) < T_L(\beta_1))] \\ &\quad = \mathbb{E}[\mathrm{e}^{-\theta T_U(\beta_2)}I(T_U(\beta_2) < T_L(\beta_1))]\mathbb{E}[\mathrm{e}^{-\theta T_L^*(W)}],\end{aligned} \tag{61}$$

and the density of W is

$$k(w) = \mu\int_0^1 \frac{1}{u}\mathrm{e}^{-\mu w/u}\mathrm{d}\mathbb{P}(F_1 < u). \tag{62}$$

Hence, from (60)

$$\mathbb{E}[e^{-\theta T_L^*(W)}] = \hat{k}(\theta+\lambda) + \int_0^\infty \mathrm{e}^{-\theta t}\frac{1}{t}\int_0^t k(w)wh^*(t-w;t)\ \mathrm{d}w\ \mathrm{d}t, \tag{63}$$

where $\hat{k}(\theta+\lambda)$ is the LT of $k(w)$. Furthermore, by using the Wald martingale one can immediately show that in the $M/M/1$ case

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{-\theta T_L(\beta_1)} & I(T_L(\beta_1) < T_U(\beta_2))] \\ & = \frac{1}{\lambda D(\theta \mid \beta_1,\beta_2)}[(\mu+\zeta_2(\theta))\mathrm{e}^{-\beta_2\zeta_1(\theta)} - (\mu+\zeta_1(\theta))\mathrm{e}^{-\beta_2\zeta_2(\theta)}],\end{aligned} \tag{64}$$

and

$$\mathbb{E}[\mathrm{e}^{-\theta T_U(\beta_2)} I(T_U(\beta_2) < T_L(\beta_1))] = \frac{\mathrm{e}^{\beta_2\zeta_2(\theta)} - \mathrm{e}^{\beta_1\zeta_1(\theta)}}{D(\theta \mid \beta_1,\beta_2)}, \tag{65}$$

where

$$\begin{aligned}\zeta_1(\theta) &= \frac{1}{2}(\lambda-\mu+\theta) + \frac{1}{2}((\lambda-\mu+\theta)^2 + 4\theta\mu)^{1/2} \\ \zeta_2(\theta) &= \frac{1}{2}(\lambda-\mu+\theta) - \frac{1}{2}((\lambda-\mu+\theta)^2 + 4\theta\mu)^{1/2}\end{aligned} \tag{66}$$

and

$$D(\theta \mid \beta_1,\beta_2) = \frac{1}{\lambda}[(\mu+\zeta_2(\theta))\mathrm{e}^{-\beta_2\zeta_1(\theta)+\beta_1\zeta_2(\theta)} - (\mu+\zeta_1(\theta))\mathrm{e}^{-\beta_2\zeta_2(\theta)+\beta_1\zeta_1(\theta)}]. \tag{67}$$

We denote the restricted LST (64) by $M_L(\theta \mid \beta_1,\beta_2)$ and the one in (65) by $M_U(\theta \mid \beta_1,\beta_2)$. We then obtain that

$$M_L^{(I)}(\theta \mid b) = \mu \int_0^b \mathrm{e}^{-\mu x} M_L(\theta \mid x, b-x)\ \mathrm{d}x. \tag{68}$$

Similarly,

$$M_U^{(I)}(\theta \mid b) = \left(\mu \int_0^b \mathrm{e}^{-\mu x} M_U(\theta \mid x, b-x)\ \mathrm{d}x + \mathrm{e}^{-\mu b} \right) \mathbb{E}[\mathrm{e}^{-\theta T_L^*(W)}]. \tag{69}$$

Also,

$$\psi_L^*(\theta \mid b) = M_L(\theta \mid b, 0) \tag{70}$$

and

$$\psi_C^*(\theta \mid b) = M_U(\theta \mid b, 0)\mathbb{E}[\mathrm{e}^{-\theta T_L^*(W)}]. \tag{71}$$

Substituting (68)-(71) in (53) we obtain the LST of the busy period for the $M/M/1$ queue under QRA.

6 The $M/G/1$ Case

In the general $M/G/1$ case, the distribution of the overshoot (47) depends on the stopping time $T_U(\beta_2)$. We develop here the joint density of $(T_U(\beta_2), R_1)$ on the set $\{T_L(\beta_1) > T_U(\beta_2)\}$. For this we have to introduce the following defective densities on $(0,\infty)$:

$$g_{\beta_2}(y;t) = \frac{d}{dy}\mathbb{P}(Y(t) \le y, T_U(\beta_2) > t) \tag{72}$$

and

$$g(y;t,\beta_1,\beta_2) = \frac{d}{dy}\mathbb{P}(Y(t) \le y, \min(T_L(\beta_1), T_U(\beta_2)) > t). \tag{73}$$

It is proved in [33] that

$$g_0(y;t) = \frac{(t-y)^+}{t} h(y;t), \tag{74}$$

where $h(y;t)$ is the density of $Y(t)$ for $y \in (0,\infty)$, i.e.,

$$h(y,t) = \sum_{n=1}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^n}{n!} f^{(n)}(y), \quad y > 0.$$

Using (74) and the renewal-type equation

$$\begin{aligned} g_{\beta_2}(y;t) = h(y;t) - 1_{(\beta_2,\beta_2+t)}(y) \Bigg[& \mathrm{e}^{-\lambda(t+\beta_2-y)} h(y; y-\beta_2) \\ & + \int_{\beta_2}^{y} h(u; u-\beta_2) g_0(y-u; t+\beta_2-u) \, \mathrm{d}u \Bigg], \quad 0 \le y \le t+\beta_2, \end{aligned} \tag{75}$$

we obtain $g_{\beta_2}(y;t)$ explicitly for $0 < y < t+\beta_2$:

$$\begin{aligned} g_{\beta_2}(y;t) = h(y;t) - 1_{(\beta_2,\beta_2+t)}(y) \Bigg[& \mathrm{e}^{-\lambda(t+\beta_2-y)} h(y; y-\beta_2) \\ & + (t+\beta_2-y) \int_0^{y-\beta_2} h(u+\beta_2; u) \frac{1}{t-u} h(y-u-\beta_2; t-u) \, \mathrm{d}u \Bigg]. \end{aligned} \tag{76}$$

The function $g(y;t,\beta_1,\beta_2)$ can now be written in terms of $g_\beta(y;t)$. Let $\delta = \beta_1 + \beta_2$. For $(t-\beta_1)^+ < y < t+\beta_2$ we have

$$\begin{aligned} g(y;t,\beta_1,\beta_2) = g_{\beta_2}(y;t) - 1_{(\beta_1,\infty)}(t) \Bigg[& \mathrm{e}^{-\lambda\beta_1} g_\delta(y; t-\beta_1) \\ & + \beta_1 \int_{\beta_1}^{y} \frac{1}{s} g_{\beta_2}(s-\beta_1; s) g_\delta(y-s+\beta_1; t-s) \, \mathrm{d}s \Bigg]. \end{aligned} \tag{77}$$

Finally, the joint density of $(T_U(\beta_2), R_1)$ on the set $\{T_L(\beta_1) > T_U(\beta_2)\}$ is given by

$$\begin{aligned} p(t,r;\beta_1,\beta_2) = 1_{(0,\beta_1]}(t) \Bigg[& \lambda \mathrm{e}^{-\lambda t} f(t+\beta_2+r) \\ & + \lambda \int_0^{t+\beta_2} g_{\beta_2}(y;t) f(t+\beta_2+r-y) \, \mathrm{d}y \Bigg] \\ & + 1_{(\beta_1,\infty)}(t) \lambda \int_{t-\beta_1}^{t+\beta_2} g(y;t,\beta_1,\beta_2) f(t+\beta_2+r-y) \, \mathrm{d}y. \end{aligned} \tag{78}$$

Now we obtain all the transforms that are the building stones of the desired LST in (53).

(i) The LST of the terminating phase.
Here $\beta_1 = b$ and $\beta_2 = 0$. According to (74),

$$\begin{aligned} \psi_L^*(\theta \mid b) &= \mathbb{E}[\mathrm{e}^{-\theta T_L(b)} I(T_L(b) < T_U(0))] \\ &= \mathrm{e}^{-(\lambda+\theta)b} + b \int_b^\infty \mathrm{e}^{-\theta t} \frac{1}{t} g_0(t-b;t) \mathrm{d}t \\ &= \mathrm{e}^{-(\lambda+\theta)b} + b^2 \mathrm{e}^{-\theta b} \int_0^\infty \frac{\mathrm{e}^{-\theta t}}{(t+b)^2} h(t; t+b) \, \mathrm{d}t. \end{aligned} \tag{79}$$

(ii) The LST of the initial phase down. This LST is given by

$$\begin{aligned} M_L^{(I)}(\theta \mid b) &= \int_0^b f(x)\mathbb{E}[\mathrm{e}^{-\theta T_L(x)} I(T_L(x) < T_U(b-x))] \\ &= \int_0^b f(x)\mathrm{e}^{-(\theta+\lambda)x}\, \mathrm{d}x + \int_0^b f(x)x\mathrm{e}^{-\theta x} \int_0^\infty \frac{\mathrm{e}^{-\theta t}}{t+x} g_{b-x}(t; t+x)\, \mathrm{d}t\, \mathrm{d}x. \end{aligned} \tag{80}$$

(iii) The LST of the initial phase up.
Here we distinguish between two cases.
Case I: $X_0 > b$. In this case $R_1 = X_0 - b$ and the length of the first cycle is $T_L^*(W_0)$, where $W_0 = (X_0 - b)F_0$. Accordingly,

$$\begin{aligned} &\mathbb{E}[\mathrm{e}^{-\theta T_L^*(W_0)} \mid X_0, F_0] = \mathrm{e}^{-(\theta+\lambda)(X_0-b)F_0} \\ &\quad + (X_0-b)F_0\mathrm{e}^{-\theta(X_0-b)F_0} \int_0^\infty \frac{\mathrm{e}^{-\theta t}}{t+(X_0-b)F_0} h^*(t; t+(X_0-b)F_0)\, \mathrm{d}t. \end{aligned} \tag{81}$$

Integrating with respect to F_0 we get

$$\begin{aligned} &\mathbb{E}[\mathrm{e}^{-\theta T_L^*(W_0)} \mid X_0\} = \frac{1}{(\theta+\lambda)(X_0-b)}(1 - \mathrm{e}^{-(\theta+\lambda)(X_0-b)}) \\ &\quad + \int_0^\infty e^{-\theta t} \left(\int_0^1 \frac{(X_0-b)u}{t+(X_0-b)u} e^{-\theta(X_0-b)u} h^*(t; t+(X_0-b)u)\, \mathrm{d}u \right) \mathrm{d}t. \end{aligned} \tag{82}$$

Case II: $X_0 \leq b$. Here the length of the initial cycle up is

$$T_U(b - X_0) + T_L^*(R_1F_1).$$

Furthermore,

$$\begin{aligned} \mathbb{E}[\mathrm{e}^{-\theta T_L^*(R_1F_1)} \mid R_1] &= \frac{1}{(\theta+\lambda)R_1}(1 - \mathrm{e}^{-(\theta+\lambda)R_1}) \\ &\quad + \int_0^\infty \mathrm{e}^{-\theta t} \left(\int_0^1 \frac{R_1u}{t+R_1u} \mathrm{e}^{-\theta R_1 u} h^*(t; t+R_1u)\, \mathrm{d}u \right) \mathrm{d}t. \end{aligned} \tag{83}$$

Hence,

$$\begin{aligned} &\mathbb{E}[e^{-\theta(T_U(b-X_1)+T_L^*(R_1F_1))} \mid X_1] \\ &= \int_0^\infty \mathrm{e}^{-\theta t} \int_0^\infty p(t, r; X_1, b-X_1) \frac{1}{(\theta+\lambda)r}(1 - \mathrm{e}^{-(\theta+\lambda)r})\, \mathrm{d}r\, \mathrm{d}t \\ &+ \int_0^\infty \mathrm{e}^{-\theta t} \int_0^\infty (p(t; r; X_1, b-X_1) \int_0^\infty \mathrm{e}^{-\theta y} \bigg(\int_0^1 \frac{ru}{y+ru} \mathrm{e}^{-\theta r u} \\ &\qquad h^*(y; y+ru)\, \mathrm{d}u \bigg) \mathrm{d}y\, \mathrm{d}r\, \mathrm{d}t. \end{aligned} \tag{84}$$

Finally, using (82) and (84) we can compute

$$\begin{aligned} M_U^{(I)}(\theta \mid b) &= \int_0^b f(x)\mathbb{E}[^{-\theta(T_U(b-x)+T_L^*(R_1F_1))} \mid X_1 = x]\, \mathrm{d}x \\ &\quad + \int_b^\infty f(x)\mathbb{E}[\mathrm{e}^{-\theta T_L^*((x-b)F_1)} \mid X_1 = x]\, \mathrm{d}x. \end{aligned} \tag{85}$$

(iv) The LST of the upper continuing cycle. This LST is obtained as

$$\begin{aligned}\psi^*_C(\theta \mid b) &= \mathbb{E}[\mathrm{e}^{-\theta(T_U(0)+T^*_L(R_1F_1))}] \\ &= \int_0^\infty e^{-\theta t} \int_0^\infty p(t,r;b,0)\frac{1}{(\theta+\lambda)r}(1-\mathrm{e}^{-(\theta+\lambda)r}))\ \mathrm{d}r\ \mathrm{d}t \\ &+ \int_0^\infty \mathrm{e}^{-\theta t} \int_0^\infty p(t,r;b,0) \int_0^\infty e^{-\theta y}\Bigg(\int_0^1 \frac{ru}{y+ru}\mathrm{e}^{-\theta ru} \\ &\qquad\qquad h^*(y;y+ru)\ \mathrm{d}u\Bigg)\ \mathrm{d}y\ \mathrm{d}r\ \mathrm{d}t.\end{aligned} \tag{86}$$

We have now determined the LST of the busy period of the $M/G/1$ queue under QRA.

7 The cycle maximum for the unrestricted $M/G/1$ queue

Let M be the maximum workload during the first cycle. In the following sections we derive explicit formulas for the distribution function of M for the $M/G/1$ queue (i) without restriction (this section), (ii) with fully restricted accessibility (Section 8) and (iii) under QRA (Section 9).

Let us start with the unrestricted case. Recall the upper first-exit time $T_U(\beta) = \inf\{t > 0 : Y(t) > \beta + t\}$, where $Y(t)$ is the $CPP(\lambda, B)$ defined in Section 5. We assume in this section that $\rho = \lambda\mathbb{E}[X] < 1$. Then $q(\beta) = \mathbb{P}(T_U(\beta) < \infty) < 1$ for all $\beta > 0$. It is known that the distribution function $F_M(x)$ of M is related to $q(\beta)$ via the identity

$$F_M(x) = \frac{B(x) - (f * q)(x)}{1 - q(x)}, \tag{87}$$

where f is the density of B and $f * q$ is the convolution of f and q. (87) appears in disguised form in [10], formula (7.67) on p. 618. It turns out that this remarkable identity can also be derived by the sample path renewal method on which Sections 4-6 are based. Using this technique, the following result on $T_L(\beta_1)$ and $T_U(\beta_2)$ was proved in [39]:

$$\mathbb{P}(T_L(\beta_1) < T_U(\beta_2)) = \frac{e^{-\lambda\beta_1} + g^{**}(\beta_1, \beta_2)}{1 + g^*(\beta_1 + \beta_2)}, \tag{88}$$

where, for $\delta = \beta_1 + \beta_2$,

$$g^*(\delta) = h^* - h^*_1(\delta)\left(e^{-\lambda\delta} + \frac{h^*_2(\delta) - e^{-\lambda\delta}h^*}{1+h^*}\right) \tag{89}$$

$$g^{**}(\beta_1, \beta_2) = h^*_2(\beta_1) - h^*_1(\beta_2)\left(e^{-\lambda\delta} + \frac{h^*_2(\delta) - e^{-\lambda\delta}h^*}{1+h^*}\right) \tag{90}$$

and h^*, $h^*_1(\beta)$ and $h^*_2(\beta)$ are defined by

$$h^* = \int_0^\infty h(t;t)\ \mathrm{d}t \tag{91}$$

$$h^*_1(\beta) = \int_0^\infty h(t+\beta;t)\ \mathrm{d}t \tag{92}$$

$$h^*_2(\beta) = \int_0^\infty h(t;t+\beta)\ \mathrm{d}t. \tag{93}$$

Later it was proved in [33] that if $\rho < 1$ the following relations hold:

$$h^* = \rho/(1-\rho) \tag{94}$$

$$h_1^*(\beta) = q(\beta)/(1-\rho) \tag{95}$$

$$h_2^*(\beta) = \frac{1}{1-\rho} - e^{-\lambda\beta}. \tag{96}$$

Thus in spite of being defined as seemingly complicated integrals of $(y,t) \mapsto h(y;t)$, actually h^* and $h_2^*(\beta)$ are a simple constant and a simple function, respectively, while $h_1^*(\beta)$ is the same as $q(\beta)$ except for a constant factor. Using (94) – (96), the formulas for $g^*(\delta)$ and $g^{**}(\beta_1, \beta_2)$ simplify:

$$g^*(\delta) = h^* - h_1^*(\delta) = \frac{1}{1-\rho}(\rho - q(\delta)), \tag{97}$$

and

$$g^{**}(\beta_1, \beta_2) = h_2^*(\beta_1) - h_1^*(\beta_2) = \frac{1}{1-\rho}(1 - q(\beta_2)) - e^{-\lambda\beta_1}. \tag{98}$$

Inserting (99) and (100) in (89) we obtain

$$\mathbb{P}(T_L(\beta_1) < T_U(\beta_2)) = \frac{1-q(\beta_2)}{1-q(\delta)}. \tag{99}$$

Finally, taking $\beta_2 = \xi - x$ and $\delta = \xi$ in (99) yields

$$\mathbb{P}(M < \xi \mid X_0 = x\} = \frac{1-q(\xi - x)}{1-q(\xi)}. \tag{100}$$

From (100) we immediately find (87) after deconditioning.

Explicit formulas for $q(\beta)$ are usually not available except in special cases, although of course there are various approximations (see [3]). Therefore, (87) is of limited value for the numerical computation of F_M. We will now present an alternative formula for $F_M(x)$ which again represents F_M in terms of $h(y;t)$ but without recourse to the results (89) – (96).

Recall from (72) and (76) the definition of the defective density $g_{\beta_2}(y;t)$ and a formula for this function. By Kendall's identity [5], the density of $T_L(\beta_1)$ on the set $\{T_L(\beta_1) < T_U(\beta_2)\}$, say $\psi_L(t;\beta_1,\beta_2)$, is given by

$$\psi_L(t;\beta_1,\beta_2) = \frac{\beta_1}{t} g_{\beta_2}(t-\beta_1;t), \quad t > \beta_1. \tag{101}$$

Thus, for $x < \xi$,

$$\begin{aligned}\mathbb{P}(M < \xi \mid X_0 = x) &= \mathrm{e}^{-\lambda x} + \int_x^\infty \psi_L(t;x,t-x)\;\mathrm{d}t \\ &= \mathrm{e}^{-\lambda x} + x\int_x^\infty \frac{1}{t} g_{\xi - x}(t-x;t)\;\mathrm{d}t.\end{aligned} \tag{102}$$

From (76) we obtain

$$\begin{aligned}&g_{\xi-x}(t-x;t) \\ &= h(t-x;t) - 1_{(\xi,\infty)}(t)\Big[h(t-x;t-\xi)\mathrm{e}^{-\lambda\xi} + \xi\int_\xi^t \frac{1}{u}h(u-\xi;u)h(t+\xi-x+u;t-u)\;\mathrm{d}u\Big].\end{aligned} \tag{103}$$

Accordingly, for $x < \xi$,

$$\mathbb{P}(M < \xi \mid X_0 = x) = I + II + III, \tag{104}$$

where

$$\begin{aligned} I &= e^{-\lambda x} + x \int_0^\infty \frac{1}{t+x} h(t; t+x)\ \mathrm{d}t \\ II &= -x e^{-\lambda\xi} \int_{\xi - x}^\infty \frac{1}{t+x} h(t; t - (\xi - x))\ \mathrm{d}t \\ III &= -x\xi \int_{\xi-x}^\infty \frac{1}{t+x} \Big(\int_\xi^{t+x} \frac{1}{u} h(u-\xi; u) h(t+\xi-u; t+x-u)\ \mathrm{d}u \Big)\ \mathrm{d}t. \end{aligned} \tag{105}$$

We now simplify the three components I – III of the distribution function of M. First, we notice that $t \mapsto \frac{x}{t} h(t-x; t)$ is the density of $T_L(x)$. Thus,

$$I = \mathrm{e}^{-\lambda x} + x \int_x^\infty \frac{1}{t} h(t-x; t)\ \mathrm{d}t = 1. \tag{106}$$

Second,

$$II = -x\mathrm{e}^{-\lambda\xi} \int_0^\infty \frac{1}{t+\xi} h(t + \xi - x; t)\ \mathrm{d}t,$$

and making the change of variable $z = \dfrac{\xi}{t+\xi}$ we get

$$II = -xe^{-\lambda\xi} \int_0^1 \frac{1}{z} h\left(\frac{\xi}{z} - x; \frac{\xi}{z} - \xi\right) \mathrm{d}z. \tag{107}$$

Third, regarding III, define

$$E(u,v) = \int_0^1 \frac{1}{z} h\left(\frac{u}{z} - u + v; \frac{u}{z} - u\right)\ \mathrm{d}z \tag{108}$$

and note that

$$\begin{aligned} E(u, \xi - x) &= \int_0^1 \frac{1}{z} h\left(\frac{u}{z} - u + \xi - x; \frac{u}{z} - u\right)\ \mathrm{d}z \\ &= \int_0^\infty \frac{1}{\omega + u} h(\omega + \xi - x; \omega)\ \mathrm{d}\omega \\ &= \int_{u-x}^\infty \frac{1}{t+x} h(t + \xi - u; t + x - u)\ \mathrm{d}t. \end{aligned} \tag{109}$$

Hence, we can write

$$\begin{aligned} III &= -\xi x \int_\xi^\infty \frac{1}{u} h(u-\xi; u) \int_{u-x}^\infty \frac{1}{t+x} h(t+\xi-u; t+x-u)\ \mathrm{d}u\ \ \mathrm{d}t \\ &= -\xi x \int_\xi^\infty \frac{1}{u} h(u-\xi; u) E(u; \xi - x)\ \mathrm{d}u. \end{aligned}$$

Summarizing, we obtain the following new formula for the distribution function of M:

$$\begin{aligned}F_M(\xi) &= \int_0^\xi f(x)\mathbb{P}(M < \xi \mid X_0 = x)\, \mathrm{d}x \\ &= F(\xi) - \mathrm{e}^{-\lambda\xi} \int_0^1 \frac{1}{z} \int_0^\xi f(x)h\left(\frac{\xi}{z} - x; \frac{\xi}{z} - \xi\right)\, \mathrm{d}x\, \mathrm{d}z \\ &- \xi \int_\xi^\infty \frac{1}{u} h(u-\xi; u) \int_0^\xi f(x)E(u; \xi - x)\, \mathrm{d}x\, \mathrm{d}u.\end{aligned} \tag{110}$$

Remark. As noted in [28], there is a close relationship between F_M and the steady-state distribution function of $V(t)$, say G_{eq}. We have

$$G_{eq}(x) = \exp\left\{-\lambda \int_x^\infty \mathbb{P}(M_c > u)\, \mathrm{d}u\right\}. \tag{111}$$

Accordingly, the steady-state probability that the queue is idle is

$$G_{eq}(0) = \exp\{-\lambda\mathbb{E}[M_c]\} \tag{112}$$

On the other hand, considering the alternating renewal process of idle and busy periods yields an alternative expression for this probability:

$$G_{eq}(0) = \frac{1/\lambda}{(1/\lambda) + \mathbb{E}[T]}, \tag{113}$$

where T is the duration of the first busy period. From (112) and (113) we get the following interesting relationship between the expected length of the busy period $\mathbb{E}[T]$ and the expected cycle maximum $\mathbb{E}[M]$, (see also [3, pp. 618, eq. (7.70)]):

$$\mathbb{E}[T] = \frac{1}{\lambda} \log(1 + \lambda\mathbb{E}[M]). \tag{114}$$

■

8 The cycle maximum under restricted accessibility

We now derive the distribution of the cycle maximum for the $M/G/1$ system under the following admission policy: all customers that arrive while the workload process is below level b receive full service, but as long as the workload is above b no new customers are admitted.

Let us determine $\mathbb{P}(M < \xi)$ for this system. First note that $\mathbb{P}(M < \xi)$ is given by (110) if $\xi \leq b$. But of course M will be larger than b if $V(t)$ jumps above b during the busy period. As in Section 4 we distinguish three types of phases that could take place during a busy period: an *initial phase*, the *continuing phase* and a *terminal phase*. The initial phase starts at the beginning of a busy period till the first time when b is exceeded. At each crossing time of $V(t)$ from above b to below b, a new phase starts; it ends when $V(t)$ either jumps again above b (then it is called a continuing phase) or hits zero (in this case it is a terminating phase). For the first phase we define $\beta_1^{(1)} = \min(X_0, b)$ and $\beta_2^{(1)} = b - \beta_1$. The first phase is initial if $T_L(\beta_1^{(1)}) > T_U(\beta_2^{(1)})$, otherwise the first phase is terminating. If it is initial, we have to consider an independent two-sided first-exit problem with $\beta_1^{(2)} = b$ and $\beta_2^{(2)} = 0$ for the second phase. If the upper boundary is reached before the lower boundary, the second phase is a continuing one, otherwise it is terminating, etc.

Notice that if $X_0 > b$ the first phase is already either continuing or terminating, while if $X_0 \le b$ and the first phase is initial (not terminating), its length is $T_U(b - X_0)$ and it ends with a jump above b having a certain overshoot. The continuing phases are independent replications of the first one; their number N has the geometric distribution

$$\mathbb{P}(N = n) = \mathbb{P}(T_L(b) < T_U(0))[\mathbb{P}(T_U(0) < T_L(b))]^{n-1}, \quad n \ge 1. \tag{115}$$

In Section 6 we have introduced the auxiliary (defective) densities $g_{\beta_2}(y;t)$, $g(y;t,\beta_1,\beta_2)$ and expressed them in terms of $h(y;t)$; cf. (76) and (77). Then we have derived the defective density $p(t,r;\beta_1,\beta_2)$ in terms of $g_{\beta_2}(y;t)$ and $g(y;t,\beta_1,\beta_2)$ in eq. (78). This latter density is now needed: Clearly, for $\xi > b$, the probability that the jump above b in the initial phase is smaller than ξ is

$$P_1(\xi) = \int_0^b f(x)\Big(\int_0^\infty \int_0^{\xi-b} p(t,r;x,b-x)\ \mathrm{d}r\ \mathrm{d}t\Big)\ \mathrm{d}x. \tag{116}$$

Similarly, the probability that the jump above b in a continuing phase is less than ξ is

$$P_c(\xi) = \int_0^\infty \int_0^{\xi-b} p(t,r;b,0)\ \mathrm{d}r\ \mathrm{d}t. \tag{117}$$

Furthermore, by (101) and (74),

$$\begin{aligned}\mathbb{P}(T_L(b) < T_U(0)) &= \mathrm{e}^{-\lambda b} + \int_b^\infty \psi_L(t;b,0)\ \mathrm{d}t \\ &= \mathrm{e}^{-\lambda b} + \int_b^\infty \frac{b}{t} g_0(t-b;t)\ \mathrm{d}t \\ &= \mathrm{e}^{-\lambda b} + b^2 \int_b^\infty \frac{1}{t^2} h(t-b;t)\ \mathrm{d}t.\end{aligned} \tag{118}$$

Finally, since the jumps above b in continuing phases are i.i.d. we obtain, for $b < \xi$,

$$\begin{aligned}\mathbb{P}(M < \xi) &= \mathbb{P}(M < b) + [P_1(\xi) + (F(\xi) - F(b))P_c(\xi)]\mathbb{P}(T_L(b) < T_U(0)) \sum_{n=0}^\infty P_c(\xi)^n \\ &= F_M(b) + P(T_L(b) < T_U(0)) \frac{P_1(\xi) + (F(\xi) - F(b))P_c(\xi)}{1 - P_c(\xi)},\end{aligned} \tag{119}$$

where $P(T_L(b) < T_U(0))$, $P_1(\xi)$ and $P_c(\xi)$ are given by (118), (116) and (117), respectively.
Remark 1. The marginal density of $T_U(\beta_2)$ on the set $\{T_L(\beta_1) > T_U(\beta_2)\}$ is

$$\begin{aligned}\psi(t;\beta_1,\beta_2) &= \lambda \mathrm{e}^{-\lambda_t} \bar{F}(\beta_2 + t) \\ &\quad + 1_{[0,\beta_1]}(t)\lambda \int_0^{\beta_2+t} g_{\beta_2}(x;t)\bar{F}(\beta_2 + t - x)\ \mathrm{d}x \\ &\quad + 1_{(\beta_1,\infty)}(t)\lambda \int_{t-\beta_1}^{t+\beta_2} g(x;t,\beta_1,\beta_2)\bar{F}(\beta_2 + t - x)\ \mathrm{d}x.\end{aligned} \tag{120}$$

■

Remark 2. In the $M/M/1$ case with $f(x) = \mu e^{-\mu x}$, $x > 0$, we have

$$p(t, r; \beta_1, \beta_2) = \mu e^{-\mu r} \psi(t; \beta_1, \beta_2). \tag{121}$$

Hence, for $\xi > b$,

$$\begin{aligned} P_c(\xi) &= (1 - e^{-\mu(\xi - b)}) \int_0^\infty \psi(t; b, 0) \, \mathrm{d}t \\ &= (1 - e^{-\mu(\xi-b)}) \mathbb{P}(T_U(0) < T_L(b)) \\ &= (1 - \mathrm{e}^{-\mu(\xi-b)})(1 - \mathbb{P}(T_L(b) < T_U(0))). \end{aligned} \tag{122}$$

Moreover,

$$P_1(\xi) = \int_0^b f(x)(1 - \mathbb{P}(T_L(x) < T_U(b - x))) \, \mathrm{d}x (1 - e^{-\mu(\xi-b)}) \tag{123}$$

and

$$\mathbb{P}(T_L(\beta_1) < T_U(\beta_2)) = \frac{\mu e^{\beta_2(\mu-\lambda)} - \lambda}{\mu e^{\beta_2(\mu-\lambda)} - \lambda e^{-\beta_1(\mu-\lambda)}}. \tag{124}$$

So in the $M/M/1$ case we obtain the distribution function F_M in closed form. ■

Remark 3. In the alternative finite dam model a customer whose service requirement would increase the current workload, say v, above level b receives only the amount $b - v$ so as to reach the capacity limit. The cycle maximum in this model variant is easily seen to have the same distribution as the cycle maximum of the unrestricted $M/G/1$ queue truncated at b. Thus $\mathbb{P}(M < x)$ is given by (87) for $x \leq b$ and

$$\mathbb{P}(M = b) = 1 - \mathbb{P}(M < b) = 1 - \frac{B(b) - (q * f)(b)}{1 - q(b)}. \tag{125}$$

■

9 Distribution of M under QRA

Under QRA, as long as $V(t)$ stays above b, each demand for service is shrunk and has the density

$$f^*(x) = \int_0^1 \frac{1}{u} g(u) f\left(\frac{x}{u}\right) \mathrm{d}u, \tag{126}$$

where g is the density of the random fraction F. Let $q^*(\beta)$ be the ruin probability corresponding to f^*, i.e.,

$$q^*(\beta) = (1 - \rho) \int_0^\infty h^*(t + \beta; t) \, \mathrm{d}t \tag{127}$$

where $h^*(y; t) = \sum_{n=1}^\infty (\mathrm{e}^{-\lambda t} (\lambda t)^n / n!) f^{*(n)}(y)$. Thus, given that the first jump above b in a continuing phase is of size $R = r$ (before shrinking) and is immediately multiplied by the random factor $F = u$, then for $\xi > b$, the conditional distribution of the maximum of $V(t) - b$ during this phase, say $\tilde{M}$, is given by

$$\mathbb{P}(\tilde{M} < \xi - b \mid R = r, F = u) = \frac{1 - q^*(\xi - b - ru)}{1 - q^*(\xi - b)}; \tag{128}$$

recall (100). Accordingly, the distribution function of $\tilde{M}$ in the first continuing phase is

$$\begin{aligned} P_1(\xi) = &\int_0^b f(x) \int_0^\infty \int_0^1 \frac{1}{u} g(u) \int_0^{\xi-b} p\left(t, \frac{r}{u}; x, b-x\right) \frac{1-q^*(\xi^*-b-ru)}{1-q^*(\xi-b)} \,\mathrm{d}r\,\mathrm{d}u\,\mathrm{d}t\,\mathrm{d}x \\ &+ (F(\xi)-F(b)) \int_0^\infty \int_0^1 \frac{1}{u} g(u) \int_0^{\xi-b} p\left(t; \frac{r}{u}; b, 0\right) \frac{1-q^*(\xi-b-ru)}{1-q^*(\xi-b)} \,\mathrm{d}r\,\mathrm{d}u\,\mathrm{d}t. \end{aligned} \tag{129}$$

For the following continuing phases, the distribution of the corresponding phase maximum minus b is

$$\tilde{P}(\xi) = \int_0^\infty \int_0^1 \frac{1}{u} g(u) \int_0^{\xi-b} p\left(t, \frac{r}{u}; b, 0\right) \frac{1-q^*(\xi-b-ru)}{1-q^*(\xi-b)} \,\mathrm{d}r\,\mathrm{d}u\,\mathrm{d}t. \tag{130}$$

Finally, the distribution function of the total cycle maximum M under QRA is, for $\xi > b$,

$$\begin{aligned} F_M(\xi) &= \mathbb{P}(M \le b) + \mathbb{P}(0 < M-b < \xi-b) \\ &= F_M(b) + \mathbb{P}(T_U(b) < T_U(0)) \frac{\tilde{P}_1(\xi)}{1-\tilde{P}(\xi)}. \end{aligned} \tag{131}$$

10 The busy period of the M/Erlang(2,μ)/1 queue under QRA

For phase-type service distributions there is an alternative way to obtain the LST of the busy period under QRA which leads to closed-form solutions. We illustrate this approach in the case of Erlang(2,μ)-distributed service requirements and uniform random fractions F_n. The approach can be extended to general phase-type distributions and general F_n.

Consider a busy period initiated by a service requirement $X_0 = x$ at time zero. We use the notation $E_x[\cdot]{=}E[\cdot \mid X_0 = x]$. Let $T = \inf\{t > 0 \mid V(t) = 0\}$ be the duration of this busy period and define the auxiliary stopping time

$$\tau = \inf\{t > 0 \mid V(t) = 0 \;\text{ or }\; V(t) \ge b\}.$$

Every service requirement consists of two successive independent exp(μ)-distributed phases. Now consider the first upcrossing of level b due to a jump caused by an arriving service requirement and define the events

$$A_{3-j} = \{\text{level } b \text{ is upcrossed by the } j\text{th phase of the jump}\}, \quad j = 1, 2$$

and the three functionals

$$\begin{aligned} \phi_*(\beta; x) &= \mathbb{E}_x[\mathrm{e}^{-\beta\tau} 1_{\{V(\tau)=0\}}], \\ \phi^*(\beta; x) &= \mathbb{E}_x[\mathrm{e}^{-\beta\tau} 1_{A_2}], \\ \phi^{**}(\beta; x) &= \mathbb{E}_x[\mathrm{e}^{-\beta\tau} 1_{A_1}], \end{aligned}$$

which are so far unknown. We need to derive these improper LSTs for arbitrary $x \in (0, b]$. We now show that these three functionals can be computed from the following three linear equations:

$$\phi_*(\beta; x) + \mathrm{e}^{-\alpha_i} \frac{\mu}{\mu+\alpha_i} \phi^*(\beta; x) + \mathrm{e}^{-\alpha_i} \left(\frac{\mu}{\mu+\alpha_i}\right)^2 \phi^{**}(\beta; x) = \mathrm{e}^{-\alpha x}, \quad x \in (0, b], \quad i = 1, 2, 3 \tag{132}$$

where the $\alpha_i = \alpha_i(\beta),\ \ i = 1, 2, 3$, are the three real roots of the polynomial equation

$$\varphi(\alpha) - \beta = \alpha - \lambda \left(1 - \left(\frac{\mu}{\mu+\alpha}\right)^2\right) - \beta. \tag{133}$$

(It is easily seen that there are exactly three distinct real roots of (133).) The proof of (132) follows from an application of the Kella-Whitt martingale [19] to the compound Poisson process $Y(t)$ (defined in Section 5) with Erlang$(2, \mu)$ jumps. According to the results of [19], the process

$$M(t) = M_{\alpha,\beta}(t) = \left(\alpha - \lambda \left(1 - \left(\frac{\mu}{\mu+\alpha}\right)^2\right) - \beta\right) \int_0^t \mathrm{e}^{-\alpha Y(u) - \beta u} \mathrm{d}u + \mathrm{e}^{-\alpha x} - \mathrm{e}^{-\alpha Y(t) - \beta t},\ \ t \geq 0$$

is a martingale for all $\beta \geq 0$ and all $\alpha > -\mu$. Consider the virtual waiting time process up to time τ for which the final jump at time τ is *not* multiplied by some F, say $\{\tilde{V}(t) : 0 \leq t \leq \tau\}$. This process clearly has the same distribution as $\{x + Y(t) : 0 \leq t \leq \tau\}$. Hence an application of the optional sampling theorem to the stopping time τ yields

$$\begin{aligned}\left(\alpha - \lambda \left(1 - \left(\frac{\mu}{\mu+\alpha}\right)^2\right) - \beta\right) \mathbb{E}\left[\int_0^\tau \mathrm{e}^{-\alpha \tilde{V}(u) - \beta u} \mathrm{d}u\right] &= -\mathrm{e}^{-\alpha x} + \mathbb{E}\left[\mathrm{e}^{-\alpha \tilde{V}(\tau) - \beta\tau}\right] \\ &= -\mathrm{e}^{-\alpha x} + \phi_*(\beta; x) + \mathrm{e}^{-\alpha b} \frac{\mu}{\mu+\alpha} \phi^*(\beta; x) + \mathrm{e}^{-\alpha b} \left(\frac{\mu}{\mu+\alpha}\right)^2 \phi^{**}(\beta; x).\end{aligned} \tag{134}$$

For the second equality in (134) we have used the fact that $\tilde{V}(\tau) - b$ and τ are conditionally independent given A_{3-j} (for $j = 1, 2$). Moreover, it follows from the lack-of-memory property of the phase lengths of the jump sizes that the excess of the jump above level b is conditionally $\exp(\mu)$-distributed, given the event A_2, while it is Erlang$(2, \mu)$-distributed, given A_1. Now if we take $\alpha = \alpha_i$, $i = 1, 2, 3$, in (134), the lefthand side becomes zero and we obtain (132).

Remark 1. One of the roots α_i is smaller than $-\mu$ so that it may seem unjustified to insert this root in (134), considering that $M(t)$ is only defined for $\alpha > -\mu$. However, the righthand side of (134) is an analytic function of α in the region $\mathbb{C}\backslash\{-\mu\}$, and the function $\ell(\alpha) = \mathbb{E}[\int_0^\tau \mathrm{e}^{-\alpha \tilde{V}(u) - \beta u} \mathrm{d}u]$ can be extended analytically to this domain. Replacing the expected value in (134) by $\ell(\alpha)$, we obtain an equation between analytic functions which is valid for $\alpha \in (0, \infty)$ and thus, by analytic continuation, holds in the entire domain $\mathbb{C}\backslash\{-\mu\}$. Therefore we can take $\alpha = \alpha_i$ in this equation also if $\alpha_i < -\mu$, proving (132). ■

Remark 2. The LST of τ is given by

$$\mathbb{E}_x[\mathrm{e}^{-\beta\tau}] = \phi_*(\beta; x) + \phi^*(\beta; x) + \phi^{**}(\beta; x). \tag{135}$$

■

We are now in a position to derive the LST of T, the length of the first period. This LST satisfies the renewal-like equation

$$\mathbb{E}_x[\mathrm{e}^{-\beta T}] = \phi_*(\beta; x) + \phi^*(\beta; x)\Upsilon(\beta)\mathbb{E}_b[\mathrm{e}^{-\beta T}] + \phi^{**}(\beta; x)\Gamma(\beta)\mathbb{E}_b[\mathrm{e}^{-\beta T}],\ \ x \in (0, b] \tag{136}$$

where

$$\Gamma(\beta) = \frac{(\lambda + \mu + \beta) - \sqrt{(\lambda + \mu + \beta)^2 - 4\lambda\mu}}{2\lambda}, \tag{137}$$

and

$$\Upsilon(\beta) = \frac{\mu}{\beta + \lambda - \lambda\Gamma(\beta)} \log\left(1 + \frac{\beta + \lambda - \lambda\Gamma(\beta)}{\mu}\right). \tag{138}$$

To prove (136), we note that the decomposition in (136) corresponds to the three possible ways in which $V(t)$ can leave $(0, b]$ at time τ:

Case 1: $V(\tau) = 0$. This case contributes the term $\phi_*(\beta; x)$ to $\mathbb{E}_x[\mathrm{e}^{-\beta T}]$.

Case 2: $V(\tau) = b$ and A_1 occurs. This means that level b is upcrossed by the second phase of some jump before 0 is reached. The overshoot above b is then of the form FU, where F is uniform on $(0, 1)$, U is $\exp(\mu)$-distributed and F and U are independent. As long as $V(t)$ stays above b after time τ, all service times are of the form FS, where S is Erlang$(2, \mu)$-distributed. A simple calculation shows that in this case the product FS is $\exp(\mu)$-distributed. It follows that the LST of the time it takes from the upcrossing of b until reaching level b again is the same as that of the busy period of a special $M/M/1$ queue in which the first customer has a service time distribution with probability density

$$-\frac{d}{dx}\mathbb{P}(FU > x) = -\frac{d}{dx}\int_0^1 \mathrm{e}^{-\mu x/y}\mathrm{d}y = \int_0^1 \frac{\mu}{y}\mathrm{e}^{-\mu x/y}\mathrm{d}y,$$

while all other service times are $\exp(\mu)$. Given $FU = v$, the conditional LST of this particular busy period at β is $\exp\{-\beta v - \lambda v(1 - \Gamma(\beta))\}$; note that Γ is the LST of the busy period of the ordinary $M/M/1$ queue. Unconditioning and using the above density of FU shows that the LST of the sojourn time above b is given by (138). Following its stay above b, the workload process restarts from b independently of the past.

Case 3: $V(\tau) = b$ and A_2 occurs. The continuation of $V(t)$ after time τ can be described similarly to Case 2. The overshoot above b is now again of the form FU, where F is uniform on $(0, 1)$ and independent of U as before, but U is now Erlang$(2, \mu)$-distributed. Therefore, $FU \sim \exp(\mu)$ in this case, and the length of the time spent above B after time τ has the same distribution as the busy period of a regular $M/M/1$ queue, i.e., its LST is given by (137). As in Case 2, after falling back to b, the process $V(t)$ continues from this level independently of the past.

These arguments yield (136).

Setting $x = b$ in (136) we find that

$$\mathbb{E}_b[\mathrm{e}^{-\beta T}] = \frac{\phi_*(\beta; b)}{1 - \phi^*(\beta; b)\Upsilon(\beta) - \phi^{**}(\beta; b)\Gamma(\beta)}. \tag{139}$$

Therefore, for $x \in (0, b]$,

$$\mathbb{E}_x[\mathrm{e}^{-\beta T}] = \phi_*(\beta; x) + \frac{[\phi^*(\beta; x)\Upsilon(\beta) + \phi^{**}(\beta; x)\Gamma(\beta)]\phi_*(\beta; b)}{1 - \phi^*(\beta; b)\Upsilon(\beta) - \phi^{**}(\beta; b)\Gamma(\beta)}. \tag{140}$$

Finally, if the service requirement X_0 initiating the busy period has the Erlang$(2, \mu)$-

distribution, we have to distinguish between the cases $X_0 \in (0, b]$ and $X_0 > b$ and obtain

$$\mathbb{E}[\mathrm{e}^{-\beta T}] = \int\limits_0^b \mathbb{E}_x[e^{-\beta T}]\mu^2 x\mathrm{e}^{-\mu x}dx$$

$$+\Big(\int\limits_b^\infty\int\limits_0^1 \mu^2 x\mathrm{e}^{-\mu x}(x-b)y\exp\{-\beta(x-b)y-\lambda(x-b)y(1-\Gamma(\beta))\}\ \mathrm{d}y\ \mathrm{d}x\Big)\mathbb{E}_b([\mathrm{e}^{-\beta T})]. \tag{141}$$

Inserting (140) in (141) we arrive at a closed-form expression for the LST of the busy period T.

Remark 3. The first integral on the lefthand side in (141) is given by

$$\int\limits_0^b \mathbb{E}_x[\mathrm{e}^{-\beta T}]\mu^2 x\mathrm{e}^{-\mu x}dx = \int\limits_0^b [\phi_*(\beta;x)+\phi^*(\beta;x)+\phi^{**}(\beta;x)]\mu^2 x\mathrm{e}^{-\mu x}dx$$

and can therefore be written as a definite integral of a rational function of polynomials and exponential functions of x, since $\phi_*(\beta;x)$, $\phi^*(\beta;x)$ and $\phi^{**}(\beta;x)$, being the solutions of the three linear equations (132), are of this type. However, the resulting explicit formula is very lengthy and not very illuminating. ■

11 Busy period of Erlang$(2,\mu)/M/1$ under QRA with fixed proportions

So far we have only considered exponential interarrival times so that all systems were Markovian. In this final section we consider the *non-Markovian* Erlang$(2,\mu)/M/1$ system, with service requirements $X_n \sim \exp(\lambda)$, under QRA. We restrict ourselves to the case of 'constant proportion cutting', i.e., $F \equiv d$ for some fixed constant $d \in (0,1)$. Our aim is to derive the LST of the length of the busy period $T = \inf\{t > 0 \mid V(t) = 0\}$. This is achieved by using a duality with a certain Markovian system, which can be analyzed using the results of Section 10. The approach was introduced and applied in related contexts in [1, 26, 29]. In a similar way we can also treat more complicated phase-type distributions, but already Erlang$(2,\mu)$ interarrival times, the most simple step beyond the assumption of Poisson arrivals, will be seen to lead to a rather intricate analysis.

A typical sample path of $\mathbf{V} = \{V(t) \mid t \geq 0\}$ is depicted in the upper part of Figure 1. The interarrival times A_n can be split in two independent $\exp(\lambda)$-distributed phases. The dots in the sample path of $\mathbf{V}$ in Figure 1 mark the initial time instants of the second phases of the A_n.

The auxiliary dual process is constructed pathwise in several steps. First, replace every positive jump of $V(t)$ by a linearly increasing piece of trajectory with slope 1 on an interval whose length is equal to the jump size. Second, replace the decreasing pieces of $V(t)$, whose slopes are -1, by negative jumps whose sizes are equal to the lengths of the decrements of the pieces. For the path in Figure 1 this transformation has been carried out in the middle part of the figure. The resulting process is called $\mathbf{A} = \{A(t) \mid t \geq 0\}$. The last step of the construction is introduced just for convenience; we define the process $\mathbf{R} = \{R(t) \mid t \geq 0\}$

by $R(t) = b - A(t)$; see the lower part of Figure 1. Obviously, $R(0) = b$ and T is the first upcrossing time of level b by $\mathbf{R}$, i.e.,

$$T = \inf\{t > 0 : R(t) > b\}.$$

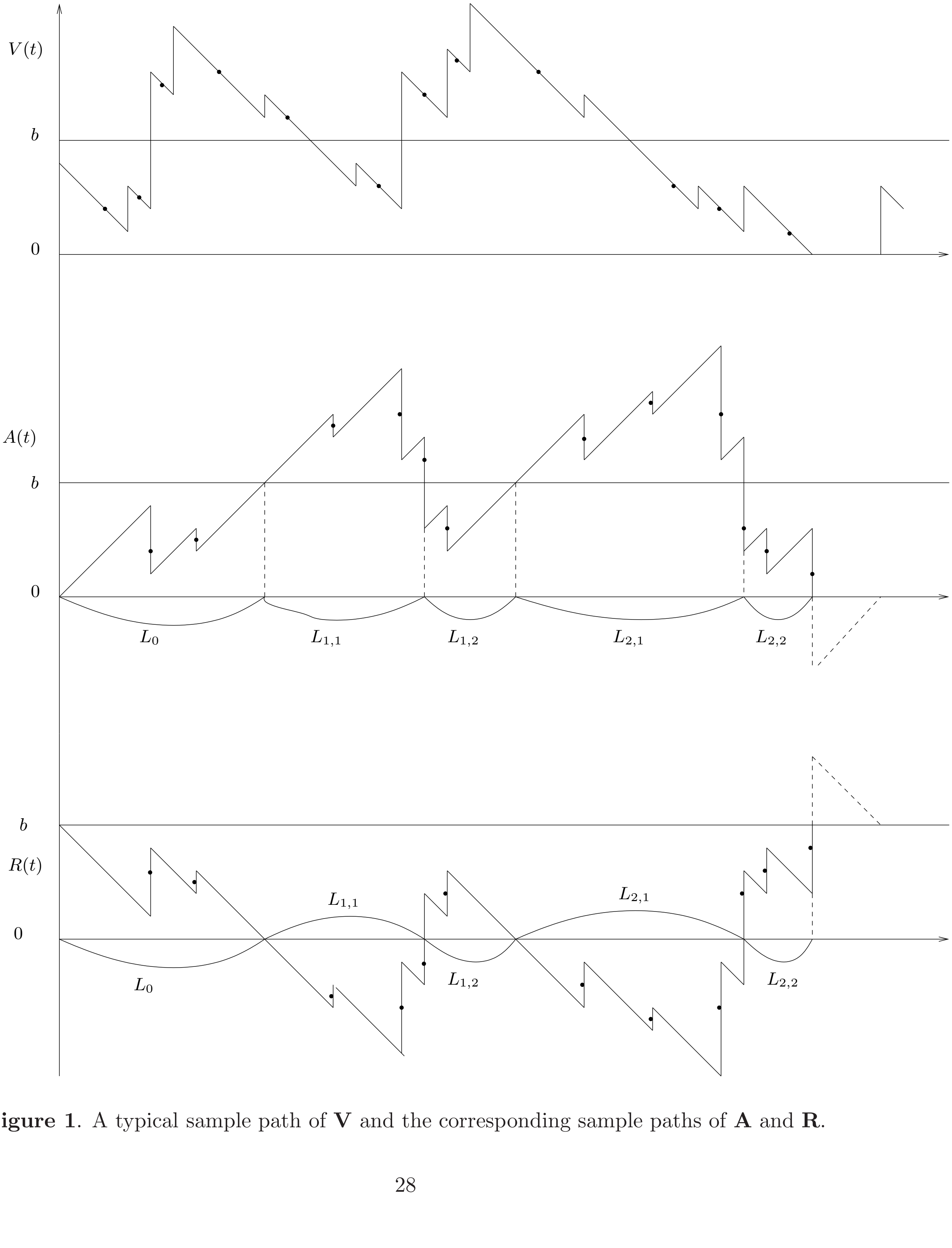

Figure 1. A typical sample path of $\mathbf{V}$ and the corresponding sample paths of $\mathbf{A}$ and $\mathbf{R}$.

The Markov process $\mathbf{R}$ is Markovian with state space $(-\infty, b]$, and it always decreases at rate -1 between Erlang$(2,\mu)$-distributed jumps. As long as $\mathbf{R} \le 0$ the jumps arrive according to a Poisson process with rate λ; when $0 < \mathbf{R} < b$ the jumps arrive according to a Poisson process with rate λ/d.

We now express T as a random sum of the following stopping times: First,

$$L_0 = \inf\{t \ge 0 \mid R(t) = 0 \ \text{ or } \ R(t) > b\}.$$

On the event $R(L_0) = 0$ we define

$$L_{1,1} = \inf\{t \ge L_0 \mid R(t) > 0\} - L_0$$

and

$$L_{1,2} = \inf\{t \ge L_0 + L_{1,1} \mid R(t) = 0 \ \text{ or } \ R(t) > b\} - L_0 - L_{1,1}.$$

On the event $R(L_0 + L_{1,1} + L_{1,2}) = 0$ we continue by defining

$$L_{2,1} = \inf\{t > L_0 + L_{1,1} + L_{1,2} \mid R(t) > 0\} - L_0 - L_{1,1} - L_{1,2}$$

and so on, until $\mathbf{R}$ upcrosses level b.

The busy period T can be expressed as the random sum

$$T = L_0 + (L_{1,1} + L_{1,2}) + ... + (L_{N,1} + L_{N,2}), \tag{142}$$

where N is the smallest index for which the righthand side of (142) is greater than b. Conditional on $N > n$, the random variables $L_0, L_{1,1} + L_{1,2}, L_{2,1} + L_{2,2}, ..., L_{n+1,1} + L_{n+1,2}$ are independent and $L_{1,1} + L_{1,2}, L_{2,1} + L_{2,2}, ..., L_{n,1} + L_{n,2}$ are also identically distributed. The regenerative structure of T makes it possible to express its LST in terms of a renewal equation.

By construction, we have $R(L_0) = 0$ or $R(L_0) > b$. Assume that $R(L_0) = 0$. Then clearly, $L_1 = L_{1,1}$ and $L_2 = L_{1,2}$ are conditionally independent given the number of the exponential phase (first or second) by which level 0 is upcrossed at time $L_0 + L_1$. $\mathbf{R}$ is negative on the time interval $(L_0, L_0 + L_1)$ and positive on $(L_0 + L_1, L_0 + L_1 + L_2)$. Next note that $\mathbb{P}(L_2 = 0) > 0$. The event $\{L_2 = 0\}$ occurs if and only if the corresponding overshoot of level 0 is also an overshoot of level b. Thus, $L_2 = 0$ implies that we have reached time T. Otherwise, if $L_2 > 0$, one of the following two events occurs: (i) If $R(L_0 + L_1 + L_2) > b$, the busy period T is equal to $L_0 + L_1 + L_2$; (ii) If $R(L_0 + L_1 + L_2) = 0$, the process $\mathbf{R}$ regenerates itself, restarting from 0.

We write $\mathbb{E}_y(\cdot) \equiv \mathbb{E}(\cdot\,|R(0) = y)$. Let $B_{3-j} = \{$level 0 is up-crossed by the jth exponential phase of the jump at time $L_0 + L_1\}$, $j = 1, 2$. We need the following functionals:

$$\eta^*(\beta) = \mathbb{E}_0(\mathrm{e}^{-\beta L_1} 1_{B_2}), \quad \eta^{**}(\beta) = \mathbb{E}_0(\mathrm{e}^{-\beta L_1} 1_{B_1}) \tag{143}$$

and

$$K_1(\beta) = \int_0^b \mu \mathrm{e}^{-\mu y} \mathbb{E}_y(\mathrm{e}^{-\beta T})\, \mathrm{d}y, \quad K_2(\beta) = \int_0^b \mu^2 y \mathrm{e}^{-\mu y}\, \mathbb{E}_y(\mathrm{e}^{-\beta T})\, \mathrm{d}y. \tag{144}$$

It turns out that we can express the LST of T in terms of the functionals in (143) and (144) via a renewal equation as follows:

$$\mathbb{E}_x(\mathrm{e}^{-\beta T}) = \phi^*(\beta;x) + \phi^{**}(\beta;x) \\ + \phi_*(\beta;x)[\eta^*(\beta)(\mathrm{e}^{-\mu b} + \mathrm{e}^{-\mu b}\mu b + K_2(\beta)) + \eta^{**}(\beta)(\mathrm{e}^{-\mu b} + K_1(\beta))]. \quad (145)$$

To prove the central formula (145), we note that for $0 < x \leq b$ we can decompose $\mathbb{E}_x(e^{-\beta T})$ into three components:

$$\mathbb{E}_x(e^{-\beta T}) = \phi^*(\beta;x) + \phi^{**}(\beta;x) + \phi_*(\beta;x)\mathbb{E}_0(e^{-\beta T}). \quad (146)$$

In (146), $\phi^*(\beta;x)$ and $\phi^{**}(\beta;x)$ represent the improper LSTs of the times until level b is first upcrossed by the first or the second exponential phase of some jump, respectively, before level 0 is reached. Similarly, $\phi_*(\beta;x)$ is the improper LST of the time until level 0 is reached before level b is upcrossed. In the latter case $\mathbf{R}$ restarts from level 0 independently of its past, which explains the product structure of the term $\phi_*(\beta;x)\mathbb{E}_0(e^{-\beta T})$. Now let us prove that

$$\mathbb{E}_0(\mathrm{e}^{-\beta T}) = \eta^*(\beta)[\mathrm{e}^{-\mu b} + \mathrm{e}^{-\mu b}\mu b + K_2(\beta)] + \eta^{**}(\beta)[\mathrm{e}^{-\mu b} + K_1(\beta)]. \quad (147)$$

To see (147), we use the conditional independence of L_1 and L_2 and distinguish between two cases.

(i) If level 0 is upcrossed by the first phase of the jump at time $L_0 + L_1$, it follows from the lack-of-memory property of that phase that the law of the overshoot is Erlang$(2, \mu)$. Then $\mathrm{e}^{-\mu b} + \mathrm{e}^{-\mu b}\mu b$ is the probability that the overshoot of level 0 is also an overshoot of level b. Otherwise, by a renewal argument, if the overshoot of level 0 is not also an overshoot of level b, the LST of the residual time until T is $K_2(\beta)$.

(ii) If level 0 is upcrossed by the second phase of the jump at time $L_0 + L_1$, it follows that the law of the overshoot is $\exp(\mu)$. Then $\mathrm{e}^{-\mu b}$ is the probability that the latter overshoot of level 0 is also an overshoot of level b. If the overshoot of level 0 is not an overshoot of level b, the LST of the residual time until T is $K_1(\beta)$.

This proves (147) and thus (145).

It remains to determine $\eta^*(\beta)$, $\eta^{**}(\beta)$, $K_1(\beta)$ and $K_2(\beta)$.

(a) Derivation of $K_1(\beta)$ and $K_2(\beta)$

Let

$$\Phi_*(\beta) = \int_0^b \mu\mathrm{e}^{-\mu y}\phi_*(\beta;y)\,\mathrm{d}y, \quad \Psi_*(\beta) = \int_0^b \mu^2 y\mathrm{e}^{-\mu y}\phi_*(\beta;y)\,\mathrm{d}y,$$

$$\Phi^*(\beta) = \int_0^b \mu\mathrm{e}^{-\mu y}\phi^*(\beta;y)\,\mathrm{d}y, \quad \Psi^*(\beta) = \int_0^b \mu^2 y\mathrm{e}^{-\mu y}\phi^*(\beta;y)\,\mathrm{d}y,$$

$$\Phi^{**}(\beta) = \int_0^b \mu\mathrm{e}^{-\mu y}\phi^{**}(\beta;y)\,\mathrm{d}y, \quad \Psi^{**}(\beta) = \int_0^b \mu^2 y\mathrm{e}^{-\mu y}\phi^{**}(\beta;y)\,\mathrm{d}y. \quad (148)$$

These six functionals can be considered to be known, because they are defined in terms of $\phi^*(\beta;y), \phi^{**}(\beta;y)$ and $\phi_*(\beta;y)$, which in turn have been computed in Section 10.

Now multiply both sides of (145) by $\mu \mathrm{e}^{-\mu x}$ and take the integral over $[0,b]$. We obtain

$$K_1(\beta) = \Phi^*(\beta) + \Phi^{**}(\beta) + \Phi_*(\beta)\eta^{**}(\beta)\mathrm{e}^{-\mu b} + K_1(\beta)\Phi_*(\beta)\eta^{**}(\beta) \\ + \Phi_*(\beta)\eta^*(\beta)[\mathrm{e}^{-\mu b} + \mathrm{e}^{-\mu b}\mu b + K_2(\beta)]$$

so that

$$K_1(\beta) = \frac{\Phi^*(\beta) + \Phi^{**}(\beta) + \Phi_*(\beta)\eta^{**}(\beta)\mathrm{e}^{-\mu b} + \Phi_*(\beta)\eta^*(\beta)[\mathrm{e}^{-\mu b} + \mathrm{e}^{-\mu b}\mu b + K_2(\beta)]}{1 - \Phi_*(\beta)\eta^{**}(\beta)}. \tag{149}$$

Similarly, multiply both sides of (145) by $\mu^2 x \mathrm{e}^{-\mu x}$, take the integral over $[0,b]$ and solve the resulting equation for $K_2(\beta)$ to get

$$K_2(\beta) = \frac{\Psi^*(\beta) + \Psi^{**}(\beta) + \Psi_*(\beta)\eta^{**}(\beta)[\mathrm{e}^{-\mu b} + K_1(\beta)] + \Psi_*(\beta)\eta^*(\beta)[\mathrm{e}^{-\mu b} + \mathrm{e}^{-\mu b}\mu b]}{1 - \Psi_*(\beta)\eta^*(\beta)}. \tag{150}$$

(149)-(150) yield the following explicit formulas:

$$\begin{aligned} K_1(\beta) = {} & \Big(1 - \Phi_*(\beta)\eta^{**}(\beta) - \Psi_*(\beta)\eta^*(\beta)\Big)^{-1} \\ & \times \Big([1 - \Psi_*(\beta)\eta^*(\beta)][\Phi^*(\beta) + \Phi^{**}(\beta) + \Phi_*(\beta)\eta^{**}(\beta)\mathrm{e}^{-\mu b} + \Phi_*(\beta)\eta^*(\beta)(\mathrm{e}^{-\mu b} + \mathrm{e}^{-\mu b}\mu b)] \\ & \quad + \Phi_*(\beta)\eta^*(\beta)[\Psi^*(\beta) + \Psi^{**}(\beta) + \Psi_*(\beta)\eta^{**}(\beta)\mathrm{e}^{-\mu b} + \Psi_*(\beta)\eta^*(\beta)(\mathrm{e}^{-\mu b} + \mathrm{e}^{-\mu b}\mu b)]\Big) \end{aligned} \tag{151}$$

and

$$\begin{aligned} K_2(\beta) = {} & \Big(1 - \Phi_*(\beta)\eta^{**}(\beta) - \Psi_*(\beta)\eta^*(\beta)\Big)^{-1} \\ & \times \Big([1 - \Phi_*(\beta)\eta^{**}(\beta)]\big(\Psi^*(\beta) + \Psi^{**}(\beta) + \Psi_*(\beta)\eta^{**}(\beta)\mathrm{e}^{-\mu b} + \Psi_*(\beta)\eta^*(\beta)[\mathrm{e}^{-\mu b} + \mathrm{e}^{-\mu b}\mu b]\big) \\ & \quad + \Psi_*(\beta)\eta^{**}(\beta)\big(\Phi^*(\beta) + \Phi^{**}(\beta) + \Phi_*(\beta)\eta^{**}(\beta)\mathrm{e}^{-\mu b} + \Phi_*(\beta)\eta^*(\beta)[\mathrm{e}^{-\mu b} + \mathrm{e}^{-\mu b}\mu b]\big)\Big). \end{aligned} \tag{152}$$

Finally we have to find $\eta^*(\beta)$ and $\eta^{**}(\beta)$.

(b) Derivation of $\eta^*(\beta)$ and $\eta^{**}(\beta)$

We start by shifting the origin to the time L_0 and define the process $\mathbf{R}_0 = \{R_0(t) \mid t \geq 0\}$ by

$$R_0(t) = R(L_0 + t), \quad t \geq 0.$$

Let us assume that $R_0(0) = R(L_0) = 0$. Clearly, $R_0(t) + t$ is a compound Poisson process with arrival rate λ/d and Erlang$(2,\mu)$-distributed jumps. For our analysis we introduce a lower boundary $-K$ and define $\sigma_K = \inf\{t > 0 \mid R_0(t) > 0 \text{ or } R_0(t) = -K\}$; later we will let K tend to ∞. The fundamental martingale identity analogous to (134) yields the relation

$$\Big(\varphi_0(\alpha) - \beta\Big)\mathbb{E}\Big(\int_0^{\sigma_K} \mathrm{e}^{-\alpha R_0(s) - \beta s}\mathrm{d}s\Big) = -1 + \mathbb{E}(\mathrm{e}^{-\alpha R_0(\sigma_K) - \beta\sigma_K}), \tag{153}$$

where

$$\varphi_0(\alpha) = \alpha - \frac{\lambda}{d}\left[1 - \left(\frac{\mu}{\mu+\alpha}\right)^2\right].$$

For all $\beta \geq 0$ the equation $\varphi_0(\alpha) = \beta$, and thus also the lefthand side of (153), has exactly three real roots $\alpha_i(\beta)$, $i = 0, 1, 2$. For $\beta > 0$ the roots can be ordered as follows: $\alpha_2(\beta) < -\mu < \alpha_1(\beta) < 0 < \alpha_0(\beta)$. For $\beta = 0$ we have $\alpha_2(0) < -\mu < \alpha_1(0) = 0 < \alpha_0(0)$. (The largest root $\alpha_0(0)$ is positive due to the stability condition $d/\lambda < 2/\mu$.) Now define the events

$$\begin{aligned}
C_1 &= \{\text{level 0 is upcrossed by the first phase of a jump of } \mathbf{R}_0 \text{ at time } \sigma_K\} \\
C_2 &= \{\text{level 0 is upcrossed by the second phase of a jump of } \mathbf{R}_0 \text{ at time } \sigma_K\} \\
C_3 &= \{\text{level } -K \text{ is hit by } \mathbf{R}_0 \text{ at time } \sigma_K\}
\end{aligned}$$

and the improper LSTs

$$\kappa_{i,K}(\beta) = \mathbb{E}(\mathrm{e}^{-\beta\sigma_K} 1_{C_i}), \quad i = 1, 2, 3.$$

By taking the roots $\alpha = \alpha_i(\beta)$ in (153) we get

$$\begin{aligned}
1 &= \mathbb{E}(\mathrm{e}^{-\alpha_i(\beta)R_0(\sigma_K)-\beta\sigma_K} 1_{C_1}) + \mathbb{E}_0(\mathrm{e}^{-\alpha_i(\beta)R_0(\sigma_K)-\beta\sigma_K} 1_{C_2}) + \mathbb{E}_0(\mathrm{e}^{-\alpha_i(\beta)R(\sigma_K)-\beta\sigma_K} 1_{C_3}) \\
&= \left(\frac{\mu}{\mu+\alpha_i(\beta)}\right)^2 \kappa_{1,K}(\beta) + \frac{\mu}{\mu+\alpha_i(\beta)}\kappa_{2,K}(\beta) + \mathrm{e}^{\alpha_i(\beta)K}\kappa_{3,K}(\beta), \quad i = 0, 1, 2.
\end{aligned} \tag{154}$$

(154) is a system of three linear equations for the three unknowns $\kappa_{i,K}(\beta)$, which can easily be solved in closed form. (The resulting formulas are however very cumbersome.) Since we have assumed the stability condition $d/\lambda < 2/\mu$, it follows that

$$\lim_{K\to\infty} \kappa_{1,K}(\beta) = \eta^{**}(\beta), \quad \lim_{K\to\infty} \kappa_{2,K}(\beta) = \eta^{*}(\beta), \quad \lim_{K\to\infty} \kappa_{3,K}(\beta) = 0.$$

If we consider the two negative roots $\alpha_1(\beta)$ and $\alpha_2(\beta)$ in (154), we see that the third term on the lefthand side converges to zero, and we arrive at the two equations

$$1 = \left(\frac{\mu}{\mu+\alpha_i(\beta)}\right)^2 \eta^{**}(\beta) + \frac{\mu}{\mu+\alpha_i(\beta)}\eta^{*}(\beta), \quad i = 1, 2. \tag{155}$$

(155) yields

$$\begin{aligned}
\eta^{*}(\beta) &= 2 - \frac{|\alpha_1(\beta)| + |\alpha_2(\beta)|}{\mu} \\
\eta^{**}(\beta) &= \frac{(\mu - |\alpha_1(\beta)|)(|\alpha_2(\beta)| - \mu)}{\mu^2}.
\end{aligned}$$

Our derivation of the LST of T is now complete.

References

[1] I. Adan, O.J. Boxma and D. Perry (2005) The $G/M/1$ queue revisited. *Mathematical Methods of Operations Research* **58**, 437-452.

[2] S. Asmussen (1998) Extreme value theory for queues via cycle maxima *Extremes* **2**, 137-168.

[3] S. Asmussen (2000) *Ruin Probabilities.* World Scientific, Singapore.

[4] S. Asmussen (2001) *Applied Probability and Queues*. 2nd ed. Springer, New York etc.

[5] K. Borovkov and Z. Burq (2001) Kendall's identity for the first crossing time revisited. *Electronic Communications in Probability* **6**, 91-94.

[6] F. Baccelli, P. Boyer and G. Hebuterne (1984) Single-server queues with impatient customers. *Advances in Applied Probability* **16**, 887-905.

[7] R. Bekker (2005) Finite-buffer queues with workload-dependent service and arrival rates. *Queueing Systems* **50**, 231-253.

[8] R. Bekker, S.C. Borst, O.J. Boxma and O. Kella (2004) Queues with workload-dependent arrival and service rates. *Queueing Systems* **46**, 537-556.

[9] Cohen, J.W. (1969) Single server queue with restricted accessibility. *Journal of Engineering Mathematics* **3**, 265-285.

[10] J.W. Cohen (1982) *The Single Server Queue*. 2nd ed. North-Holland, Amsterdam.

[11] D. J. Daley (1964) Single server queueing system with uniformly limited queueing time. *Journal of the Australian Mathematical Society* **4**, 489-505.

[12] A. Elwalid and D. Mitra (1991) Analysis and design of rate-based congestion control of high-speed networks I: stochastic fluid models, access regulation. *Queueing Systems* **9**, 29-64.

[13] A. Elwalid and D. Mitra (1994) Statistical multiplexing with loss priorities in rate-based congestion control of high-speed networks. Analysis and design of rate-based congestion control of high-speed networks. *IEEE Transactions on Communications* **42**, 2989-3002.

[14] O. Garnett, A. Mandelbaum and M. Reiman (2002) Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4**, 208-227.

[15] B. Gavish and P.J. Schweitzer (1977) The Markovian queue with bounded waiting time. *Management Science* **23**, 1349-1357.

[16] B.V. Gnedenko and I.N. Kovalenko (1989) *Introduction to Queueing Theory*, 2nd ed. Birkhäuser, Basel.

[17] G. Hooghiemstra (1987) A path construction for the virtual waiting time of an $M/G/1$ queue. *Statistica Neerlandica* **41**, 175-181.

[18] P. Hokstad (1979) A single server queue with constant service time and restricted accessibility. *Management Science* **25**, 205-208.

[19] O. Kella and W. Whitt (1992) Useful martingales for storage systems with Lévy input. *Journal of Applied Probability* **29**, 396-403.

[20] C. Knessl, B.J. Matkowsky, Z. Schuss and C. Tier (1987) Busy period distribution of state-dependent queues. *Queueing Systems* **2**, 285-305.

[21] G. Koole and A. Mandelbaum (2002) Queueing models of call-centers: an introduction. *Annals of Operations Research* **113**, 41-59.

[22] L.Q. Liu and V.G. Kulkarni (2006) Explicit solutions for the steady-state distributions in $M/PH/1$ queues with workload-dependent balking. *Queueing Systems* **52**, 251-260.

[23] J. Loris-Teghem (1972) On the waiting time distribution in a generalized queueing system with uniformly bounded sojourn time. *Journal of Applied Probability* **9**, 642-649.

[24] M. Mandjes, D. Mitra and W. Scheinhardt (2002) A simple model of network access: feedback adaptation of rates and admission control. *Proceedings of Infocom* 2002, 3-12.

[25] D. Perry and S. Asmussen (1995) Rejection rules in the $M/G/1$ queue. *Queueing Systems* **19**, 105-130.

[26] D. Perry, W. Stadje and S. Zacks (2000) Busy period analysis of $M/G/1$ and $G/M/1$ type queues with restricted accessibility. *Operations Research Letters*, **27**, 163-174.

[27] D. Perry, W. Stadje and S. Zacks (2001) The $M/G/1$ queue with finite workload capacity. *Queueing Systems* **39**, 7-22.

[28] D. Perry, W. Stadje and S. Zacks (2002) Hitting and ruin probabilities for compound Poisson processes and the cycle maximum for $M/G/1$. *Stochastic Models* **18**, 553-564.

[29] D. Perry and W. Stadje (2003) Duality of dams via mountain processes. *Operations Research Letters* **31**, 451-458.

[30] K. Ramanan and A. Weiss (1997) Sharing bandwidth in ATM. In: *Proceedings of the Allerton Conference*, 732-740.

[31] P.B.M. Roes (1970) The finite dam. *Journal of Applied Probability* **7**, 316-325 and 599-616.

[32] P.B.M. Roes (1970) The finite dam with discrete additive input. *Journal of Engineering Mathematics* **6**, 37-45.

[33] W. Stadje and S. Zacks (2003) Upper first-exit times of compound Poisson processes revisited. *Probability in the Engineering and Informational Sciences* **17**, 459-465.

[34] R.E. Stanford (1979) Reneging phenomena in single channel queues. *Mathematics of Operations Research* **4**, 162-178.

[35] D. Stoyan (1983) *Comparison Method for Queues and other Stochastic Models*. Wiley, New York.

[36] A.R. Swensen (1986) On a $GI/M/c$ queue with bounded waiting times. *Operations Research* **34**, 895-908.

[37] W. Whitt (1981) Comparing counting processes and queues. *Advances in Applied Probability* **13**, 207-220.

[38] W. Whitt (2005) Engineering solution of a basic call-center model. *Management Science* **51**, 21-235.

[39] S. Zacks, D. Perry, D. Bshouty and S. Bar-Lev (1999) Distribution of stopping times for compound Poisson processes with positive jumps and linear boundaries. *Stochastic Models* **15**, 89-101.