An M/G/1 queue with adaptable service speed

An M/G/1 queue with adaptable service speed*

R. Bekker and O.J. Boxma

Department of Mathematics & Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

and

CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

We consider a queueing system where feedback information about the level of congestion is given right after arrival instants. When the amount of work right after arrival is at most (respectively, larger than) K, then the server works at speed r_1 (respectively, r_2) until the next arrival instant. We derive the distribution of the workload right after and right before arrivals, as well as in steady state. In addition, we consider the generalization to the N-step service rule.

1 Introduction

The queueing literature contains many studies about queues with workload-dependent service speeds. In those studies it is usually assumed that the speed of the server is continuously adapted over time based on the buffer content. In many practical situations, though, service speed adaptations are only made at particular points in time, like arrival epochs. For example, feedback information about the buffer state may only be available at such epochs. Furthermore, continuously changing the service speed may come with certain costs.

In this paper, we consider a single-server queue with adaptable service speed based on the amount of work right after customer arrivals. In between arrivals, the service speed is held fixed and may not be changed until the next customer arrival. The main aim of this paper is to find the (Laplace Stieltjes Transform of the) distribution of the steady-state workload embedded at epochs immediately after arrivals, and the steady-state workload distribution at arbitrary epochs.

Related literature

Models with continuously adaptable service speed originate from the study of dams and storage processes. There exists a rich body of literature on dams and storage systems going

*The first author was financially supported by a research grant from Philips Research. The research of the second author fits into the BRICKS project. Part of the research was conducted in the framework of Euro-NGI.

back to the 1960's, see e.g. [10, 13]. Queueing systems with workload-dependent service speeds can also be found in, e.g., [1, 3, 8, 11]. Furthermore, in [7, 10] and [8], p. 555-556, the authors consider a queueing system with a two-stage service rule: If the workload is less than K, then the service speed equals r_1, whereas the service speed equals r_2 when the workload exceeds K. Using an elegant technique for the convolution of two Laplace Stieltjes Transforms (LST), they determine the steady-state workload distribution. In this paper, we apply a similar method to obtain the LST of the workload at embedded epochs for the M/G/1 queue with service speeds only being changed at customer arrivals.
A related branch of literature considers queueing systems where the service speed depends not only on the buffer content, but also on the stage of the system. In particular, an (m, M) control rule prescribes to switch from stage 1 to stage 2 at an upcrossing of the workload of level M (and the stage is 1) and to switch back from stage 2 to stage 1 at a downcrossing of m (and the stage is 2), see also e.g. [2, 12, 15]. The control of the service speed may be realized by letting r_i be the service speed in stage i, $i = 1, 2$. In such control systems, usually costs are imposed including, e.g., holding costs and switchover costs. In [15], the long-run average costs per unit time for the (m, M)-policy are determined. Of special interest is the case when $m = 0$ which is commonly referred to as a D-policy (that is $(m, M) = (0, D)$). In [14], the author shows that the D-policy is average-cost optimal under the assumption that the workload can only be controlled at arrival epochs. In [9], the average-cost optimality of D-policies is rigorously proved in a more general setting.

Model description
We consider an M/G/1 queueing system where feedback information about the level of congestion is available right after arrival instants. The customers arrive to the system according to a Poisson process with rate λ. Let A_n, $n = 1, 2, \ldots$, denote the time between the arrival instants of customers n and $n+1$. Also, denote by B_n, $n = 1, 2, \ldots$, the service requirement of customer n. We assume that $B_1, B_2, \ldots$ are i.i.d. copies of the generic random variable B with distribution $B(\cdot)$, mean β, and LST $\beta(\cdot)$. We also assume that the sequences of interarrival intervals and service requirements are independent.
When the amount of work right after an arrival instant equals x, the server works at constant speed $r(x)$ until the next customer arrival. Note that the service speed is thus only changed at discrete points in time. In this paper, we specifically consider the case of a two-step service speed function: If the amount of work right after an arrival is smaller than (or equal to) a finite number K, then the server starts to work at speed r_1, whereas the service speed equals r_2 if the workload is larger than K. Later, we also consider the generalization to an N-step service-speed function (see Subsection 5.3).
Define $\rho_i := \lambda\beta/r_i$, $i = 1, 2$. Throughout, we assume that the system is stable, i.e., $\rho_2 < 1$. Let W_n and S_n be the workload just before, respectively right after, the arrival instant of customer n. We denote by W and S the steady-state random variables corresponding to W_n and S_n. We have the following recursion relation:

$$S_{n+1} = (S_n - r(S_n)A_n)^+ + B_{n+1}, \tag{1}$$

where $x^+ = \max(x, 0)$. Because of the trivial relation $S_n = W_n + B_n$, one also has $W_{n+1} = (S_n - r(S_n)A_n)^+$.
In queueing systems where the server always works at unit speed when there is any work in the system, W corresponds to a waiting time and S represents a customer's sojourn time. This equivalence no longer holds when the service speed varies with the amount of

work present. For convenience, however, we often refer to W and S as the waiting and sojourn time, respectively.

Goal and organization

The main aim of this paper is to find the distribution (and LST) of S, and then also of W. It should be observed that, due to PASTA, the distribution of W also equals the steady-state workload distribution.

The paper is organized as follows. In Section 2 we derive two distinct equations for the LST of S and sketch a four-step procedure to determine its distribution. The first step of this procedure does not depend on the distribution of the service requirement and is analyzed in detail in Section 2. We give steps two to four in Section 3 in case the service requirements follow an exponential distribution. It turns out that the density of S is then a weighted combination of two exponentials for $x \leq K$, and is purely exponential for $x > K$. The M/M/1 case gives much insight into the structure of the solution for more general cases, like the M/G/1 case, which is addressed in Section 4. For expository reasons, we have chosen to treat these cases separately instead of all in one. Special cases and the extension to the N-step service rule are discussed in Section 5.

2 Sojourn times: Equations and general procedure

In this section, we first derive equations to determine the LST of S in case of the two-step service speed function. Secondly, we outline a four-step procedure to find the LST and distribution of S from the constructed equations, and describe the first step in detail.

For convenience, we recall the definition of the two-step service rule:

$$r(x) = \begin{cases} r_1, & \text{for } 0 < x \leq K, \\ r_2, & \text{for } x > K. \end{cases}$$

Denote the LST of S by

$$\phi(\omega) := \int_0^\infty e^{-\omega x} \mathrm{d}\mathbb{P}(S < x). \tag{2}$$

Also, define, for $i = 1, 2$ and $\rho_i \neq 1$,

$$F_i(\omega) := (1 - \rho_i) \frac{r_i \omega \beta(\omega)}{\omega r_i - \lambda + \lambda \beta(\omega)}. \tag{3}$$

Observe that $F_i(\omega)$ corresponds to the LST of the sojourn time in an M/G/1 queue with service speed r_i, $i = 1, 2$.

The equations for $\phi(\omega)$ are summarized in the following lemma:

Lemma 2.1. *$\phi(\omega)$ satisfies the following two equations, for* $\operatorname{Re} \omega \geq 0$,

$$\begin{aligned} \phi(\omega) &= F_2(\omega) \frac{W(0)}{1 - \rho_2} \\ &+ F_2(\omega) \frac{\lambda(\frac{r_1}{r_2} - 1)}{(\omega r_1 - \lambda)(1 - \rho_2)} \left[\int_0^K e^{-\omega x} \mathrm{d}\mathbb{P}(S < x) - \int_0^K e^{-\frac{\lambda}{r_1} x} \mathrm{d}\mathbb{P}(S < x) \right], \end{aligned} \tag{4}$$

with $W(0) := \mathbb{P}(W=0)$. *Also,*

$$\begin{aligned}\phi(\omega) &= F_1(\omega)\frac{W(0)}{1-\rho_1} \\ &+ F_1(\omega)\frac{\lambda(1-\frac{r_2}{r_1})}{(\omega r_2-\lambda)(1-\rho_1)}\left[\int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S<x) - \int_K^\infty e^{-\omega x}\mathrm{d}\mathbb{P}(S<x)\right].\end{aligned} \tag{5}$$

Proof. It follows after some straightforward calculations that, for $\omega \neq \lambda/r_1, \lambda/r_2$,

$$\begin{aligned}\mathbb{E}\left[e^{-\omega(S_n-r(S_n)A_n)^+}|S_n=x\right] &= e^{-\omega x}\lambda\int_0^{x/r(x)} e^{(\omega r(x)-\lambda)y}\mathrm{d}y + e^{-\lambda x/r(x)} \\ &= \frac{\omega r(x)}{\omega r(x)-\lambda}e^{-\frac{\lambda}{r(x)}x} - \frac{\lambda}{\omega r(x)-\lambda}e^{-\omega x}.\end{aligned} \tag{6}$$

Using the recursion (1), conditioning on S_n, and applying the above, yields

$$\begin{aligned}\mathbb{E}\left[e^{-\omega S_{n+1}}\right] &= \int_0^\infty \mathbb{E}\left[e^{-\omega S_{n+1}}|S_n=x\right]\mathrm{d}\mathbb{P}(S_n<x) \\ &= \beta(\omega)\Bigg[\frac{\omega r_1}{\omega r_1-\lambda}\int_0^K e^{-\frac{\lambda}{r_1}x}\mathrm{d}\mathbb{P}(S_n<x) - \frac{\lambda}{\omega r_1-\lambda}\int_0^K e^{-\omega x}\mathrm{d}\mathbb{P}(S_n<x) \\ &\quad +\frac{\omega r_2}{\omega r_2-\lambda}\int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S_n<x) - \frac{\lambda}{\omega r_2-\lambda}\int_K^\infty e^{-\omega x}\mathrm{d}\mathbb{P}(S_n<x)\Bigg].\end{aligned} \tag{7}$$

To analyze the steady-state behavior of S_n, we let $n \to \infty$. Furthermore, combining (2) and (7), in addition to some basic manipulations, we may obtain two alternative equations for $\phi(\omega)$: First,

$$\begin{aligned}\phi(\omega) &= \frac{F_2(\omega)}{(1-\rho_2)(\omega r_1-\lambda)}\Bigg[\frac{r_1}{r_2}(\omega r_2-\lambda)\int_0^K e^{-\frac{\lambda}{r_1}x}\mathrm{d}\mathbb{P}(S<x) \\ &\quad +\lambda(\frac{r_1}{r_2}-1)\int_0^K e^{-\omega x}\mathrm{d}\mathbb{P}(S<x) + (\omega r_1-\lambda)\int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S<x)\Bigg],\end{aligned} \tag{8}$$

and second,

$$\begin{aligned}\phi(\omega) &= \frac{F_1(\omega)}{(1-\rho_1)(\omega r_2-\lambda)}\Bigg[(\omega r_2-\lambda)\int_0^K e^{-\frac{\lambda}{r_1}x}\mathrm{d}\mathbb{P}(S<x) \\ &\quad -\lambda(1-\frac{r_2}{r_1})\int_K^\infty e^{-\omega x}\mathrm{d}\mathbb{P}(S<x) + \frac{r_2}{r_1}(\omega r_1-\lambda)\int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S<x)\Bigg].\end{aligned} \tag{9}$$

Now, Equations (4) and (5) follow from (8) and (9), respectively, and from the observation that

$$W(0) = \int_0^K e^{-\frac{\lambda}{r_1}x}\mathrm{d}\mathbb{P}(S<x) + \int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S<x). \tag{10}$$

This completes the proof. □

Determining $\phi(\omega)$ from Equations (4) and (5) involves the more complicated part. We introduce a four-step procedure to determine the distribution of S. Below, we sketch each of the four steps. Because Step 1 is the only step that does not depend on the service requirement distribution, we analyze it in detail at the end of this section. Steps 2–4 are

carried out in Section 3 in case the distribution of the service requirements is exponential. The general M/G/1 case is considered in Section 4. The procedure builds upon techniques applied in [7, 10] and [8], p. 556. It starts from the observation that a serious complication in determining $\phi(\omega)$ from (4) and (5) is that both equations involve the incomplete LST of S.
The basic algorithm to obtain $\mathbb{P}(S < x)$ is as follows:

Step 1 Rewrite Equation (5) such that the second term of (5) can be interpreted as the sum of (i) the LST of the convolution of $F_1(\cdot)$ with an exponential term, and (ii) a transform that only has points of increase on (K, ∞).

Step 2 Apply Laplace inversion to the reformulated Equation (5) resulting from Step 1, to determine $\mathbb{P}(S < x)$ for $x \in (0, K]$.

Step 3 By Step 2, we may now calculate $\int_0^K e^{-\omega x} \mathrm{d}\mathbb{P}(S < x)$. Substitution in (4) then directly provides $\phi(\omega)$. Applying Laplace inversion again, we determine $\mathbb{P}(S < x)$ for $x > K$.

Step 4 The remaining constants may be found by normalization.

The remainder of this section is devoted to the description of Step 1.

Step 1: *Rewriting (5)*
In this part, when considering the sojourn time of customer $n+1$, we distinguish between two cases: (i) $S_n \leq K$, and (ii) $S_n > K$. If $S_{n+1} \leq K$, this imposes for case (ii) that a downcrossing of level K occurs between the arrival instants of customers n and $n+1$. However, the residual interarrival time at a downcrossing of K is still exponential. Consequently, given a downcrossing of level K between the arrival epochs of customers n and $n+1$, the precise distribution of S_n on (K, ∞) does not affect the distribution of $S_{n+1} \leq K$, because W_{n+1} is simply distributed as $(K - r_2 A_n)^+$. The aim of this first step is to show that the second part of Equation (5) corresponds to case (ii) and to apply the intuitive arguments above in reformulating (5).
Denote by $I(\cdot)$ the indicator function. Using (6), we get

$$\begin{aligned}
&\mathbb{E}\left[e^{-\omega(S_n - r(S_n)A_n)^+} I(S_n > K)\right] - \int_K^\infty e^{-\frac{\lambda}{r_2}x} \mathrm{d}\mathbb{P}(S_n < x) \\
&\quad = \frac{\lambda}{\omega r_2 - \lambda}\left[\int_K^\infty e^{-\frac{\lambda}{r_2}x} \mathrm{d}\mathbb{P}(S_n < x) - \int_K^\infty e^{-\omega x} \mathrm{d}\mathbb{P}(S_n < x)\right].
\end{aligned}$$

Observe that the right-hand side (rhs) corresponds to the final part of the second term in

(5). However, by conditioning on S_n, we may also rewrite this expression as

$$
\begin{aligned}
&\mathbb{E}\left[e^{-\omega(S_n - r(S_n)A_n)^+} I(S_n > K)\right] - \int_K^\infty e^{-\frac{\lambda}{r_2}x} \mathrm{d}\mathbb{P}(S_n < x) \\
&= \int_K^\infty e^{-\omega(x - r_2 A_n)^+} \mathrm{d}\mathbb{P}(S_n < x) - \int_K^\infty e^{-\frac{\lambda}{r_2}x} \mathrm{d}\mathbb{P}(S_n < x) \\
&= \int_K^\infty e^{-\omega(x - r_2 A_n)} I(A_n \leq (x-K)/r_2) \mathrm{d}\mathbb{P}(S_n < x) \\
&\qquad + \int_K^\infty \int_{(x-K)/r_2}^{x/r_2} \lambda e^{-\lambda y} e^{-\omega(x - r_2 y)} \mathrm{d}y \mathrm{d}\mathbb{P}(S_n < x) \\
&= \mathbb{E}\left[e^{-\omega(S_n - r_2 A_n)} I(S_n - r_2 A_n > K)\right] \\
&\qquad + \frac{\lambda}{\omega r_2 - \lambda}\left(1 - e^{-\omega K + \frac{\lambda}{r_2}K}\right) \int_K^\infty e^{-\frac{\lambda}{r_2}x} \mathrm{d}\mathbb{P}(S_n < x).
\end{aligned}
$$

Letting $n \to \infty$ and combining the above, Equation (5) reads,

$$
\begin{aligned}
\phi(\omega) &= F_1(\omega)\frac{W(0)}{1-\rho_1} + F_1(\omega)\frac{(1-\frac{r_2}{r_1})}{1-\rho_1}\Big\{\mathbb{E}\left[e^{-\omega(S - r_2 A)} I(S - r_2 A > K)\right] \\
&\qquad + \frac{\lambda}{\omega r_2 - \lambda}\left(1 - e^{-\omega K + \frac{\lambda}{r_2}K}\right) \int_K^\infty e^{-\frac{\lambda}{r_2}x} \mathrm{d}\mathbb{P}(S < x)\Big\}. \qquad (11)
\end{aligned}
$$

The second and third term on the rhs of (11) directly correspond to the intuitive observations made above. The first one provides the LST of W when $W > K$. The second one involves the LST of $K - r_2 A$ (with A a generic interarrival time) multiplied by a constant (see Section 4 for an interpretation).

3 Exponential service requirements

In this section, we assume that $B(x) = 1 - e^{-\mu x}$, i.e., the service requirements are exponentially distributed with mean $1/\mu$. Applying the procedure described in Section 2, we explicitly determine the steady-state "sojourn time" distribution. We have chosen to treat the M/M/1 case first, because the structure of the density of S is here readily exposed, yielding insight into the solution for the M/G/1 case. Moreover, the solutions reduce to nice analytical expressions in that case.
Because the interpretation of Step 1 is valid independently of $B(\cdot)$, the starting point of the algorithm is Equation (11).

Step 2: *Sojourn time density on* $(0, K]$
Using the construction of Step 1, we apply Laplace inversion to determine the density $f_S(x)$ of S for $0 < x \leq K$. In the exponential case, we easily obtain for the first transform in (11),

$$F_1(\omega) = (1-\rho_1)\frac{r_1\mu}{r_1(\omega+\mu) - \lambda}.$$

Laplace inversion provides the familiar M/M/1 term for queues with constant service speed r_1,

$$s_1(x) = \mu(1-\rho_1)e^{(\frac{\lambda}{r_1} - \mu)x}, \qquad \text{for } x > 0,$$

where $s_1(\cdot)$ denotes the density of a random variable with LST $F_1(\cdot)$.
The inversion of the second transform in (11) is based on an observation made in [7, 8, 10].
First, consider

$$F_1(\omega)\mathbb{E}\left[e^{-\omega(S-r_2A)}I(S-r_2A>K)\right].$$

This term involves a product of two LST, corresponding to the sum of a random variable with mass on $[0,\infty)$, and one with mass on $[K,\infty)$. Hence, that sum has no mass on $[0,K]$.
Second, consider

$$F_1(\omega)\frac{\lambda}{\omega r_2-\lambda}\left(1-e^{-\omega K+\frac{\lambda}{r_2}K}\right). \tag{12}$$

It is readily checked that the latter part, $\frac{\lambda}{\omega r_2-\lambda}\left(1-e^{-\omega K+\frac{\lambda}{r_2}K}\right)$, is the Laplace Transform of the function

$$f(x)=\begin{cases}\frac{\lambda}{r_2}e^{\frac{\lambda}{r_2}x}, & \text{for } 0<x\leq K,\\ 0, & \text{for } x>K.\end{cases} \tag{13}$$

Thus, (12) represents the convolution of $s_1(\cdot)$ with $f(\cdot)$. By applying (11) and combining the above, we obtain after lengthy calculations the following "sojourn time" density $f_S(x)$, for $0<x\leq K$,

$$\begin{aligned} f_S(x) &= s_1(x)\frac{W(0)}{1-\rho_1}+\frac{1-\frac{r_2}{r_1}}{1-\rho_1}\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S<y)\int_0^x s_1(y)\frac{\lambda}{r_2}e^{\frac{\lambda}{r_2}(x-y)}\mathrm{d}y \\ &= Q_1e^{(\frac{\lambda}{r_1}-\mu)x}+Q_2e^{\frac{\lambda}{r_2}x}, \end{aligned} \tag{14}$$

with

$$Q_1 = \mu\int_0^K e^{-\frac{\lambda}{r_1}y}\mathrm{d}\mathbb{P}(S<y)+\frac{r_1r_2\mu^2}{\lambda(r_1-r_2)+r_1r_2\mu}\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S<y), \tag{15}$$

$$Q_2 = \frac{\lambda\mu(r_1-r_2)}{\lambda(r_1-r_2)+r_1r_2\mu}\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S<y). \tag{16}$$

Because we have determined the density of S on $(0,K]$ up to some constants, this concludes Step 2.

Step 3: *Sojourn time density on* (K,∞)
In this step, we first determine $\phi(\omega)$ using (4) and then apply Laplace inversion once more to obtain the density of S on (K,∞). From the final result of Step 2, we deduce,

$$\int_0^K e^{-\omega x}\mathrm{d}\mathbb{P}(S<x)=\frac{Q_1}{\omega+\mu-\lambda/r_1}(1-e^{(\frac{\lambda}{r_1}-\mu-\omega)K})+\frac{Q_2}{\omega-\lambda/r_2}(1-e^{(\frac{\lambda}{r_2}-\omega)K}). \tag{17}$$

Substitution in (4) then immediately yields $\phi(\omega)$.
Next, to obtain $f_S(x)$ for $x>K$, we invert $\phi(\omega)$ on the corresponding interval. Similar to $F_1(\omega)$ in Step 2, we have

$$F_2(\omega)=(1-\rho_2)\frac{r_2\mu}{r_2(\omega+\mu)-\lambda}.$$

Laplace inversion provides the expression of an M/M/1 queue with service speed r_2:

$$s_2(x)=\mu(1-\rho_2)e^{(\frac{\lambda}{r_2}-\mu)x}, \qquad \text{for } x>0,$$

where $s_2(\cdot)$ represents the density of a random variable with LST $F_2(\cdot)$.
By (17), it follows that the second term of Equation (4) constitutes a Laplace transform having four poles. We observe that the zero in the denominator of $\lambda/(\omega r_1 - \lambda)$ is a removable zero. The expression in (17) is the LST of a density on $(0, K]$. Hence, the only pole contributing on (K, ∞) is the zero in the denominator of $F_2(\omega)$, that is, $\eta = \lambda/r_2 - \mu$. Since the first term of (4) provides the same pole, we immediately deduce that

$$f_S(x) = Q_3 e^{(\frac{\lambda}{r_2}-\mu)x}, \qquad \text{for } \; x > K. \tag{18}$$

We note that the terms with removable singularities in $\lambda/(\omega r_1 - \lambda)$ and (17) do affect the constant Q_3. However, Q_3 is determined in Step 4 using the expressions for Q_1, Q_2, and the normalizing condition, and there is thus no need to specify Q_3 any further.

Step 4: *Determination of the constants*
In this final step, we use the normalizing condition $\int_0^\infty \mathrm{d}\mathbb{P}(S < x) = 1$ to determine the constants Q_1, Q_2, and Q_3. In particular, combining normalization with (15) and (16) we obtain a set of three equations with the above three unknowns (hence, there is indeed no need to give Q_3 explicitly in Step 3).
Substituting (18) in (16) and calculating the integral yields

$$Q_2 = Q_3 \frac{\lambda(r_1 - r_2)}{\lambda(r_1 - r_2) + r_1 r_2 \mu} e^{-\mu K}. \tag{19}$$

Also, substitution of both (14) and (18) in (15) and performing the integrations, yield, for $r_1 \neq r_2$,

$$Q_1 = Q_1(1 - e^{-\mu K}) + Q_2 \frac{r_1 r_2 \mu}{\lambda(r_1 - r_2)}(e^{(\frac{\lambda}{r_2}-\frac{\lambda}{r_1})K} - 1) + Q_3 \frac{r_1 r_2 \mu}{\lambda(r_1 - r_2) + r_1 r_2 \mu} e^{-\mu K}.$$

Consequently, using the expression for Q_2 in (19) in addition to some rewriting, we express Q_1 in terms of Q_3 as

$$Q_1 = Q_3 \frac{r_1 r_2 \mu}{\lambda(r_1 - r_2) + r_1 r_2 \mu} e^{(\frac{\lambda}{r_2}-\frac{\lambda}{r_1})K}. \tag{20}$$

We obtain an additional equation from the normalizing condition $\int_0^\infty f_S(x)\mathrm{d}x = 1$. Using the densities of (14) and (18) and determining the integrals yields (for $\lambda \neq r_1\mu$, with an obvious modification when $\lambda = r_1\mu$):

$$\frac{Q_1 r_1}{\lambda - r_1 \mu}(e^{(\frac{\lambda}{r_1}-\mu)K} - 1) + \frac{Q_2 r_2}{\lambda}(e^{\frac{\lambda}{r_2}K} - 1) + \frac{Q_3 r_2}{\lambda - r_2 \mu} e^{(\frac{\lambda}{r_2}-\mu)K} = 1.$$

Now, substituting (20) and (19) in the above in addition to some manipulations, gives

$$\begin{aligned} Q_3 \;\; = \;\; & \Bigg[\left(\frac{r_2}{\lambda - r_1\mu} - \frac{r_2}{\lambda - r_2\mu}\right) e^{(\frac{\lambda}{r_2}-\mu)K} - \frac{r_1^2 r_2 \mu}{(\lambda(r_1 - r_2) + r_1 r_2 \mu)(\lambda - r_1 \mu)} e^{(\frac{\lambda}{r_2}-\frac{\lambda}{r_1})K} \\ & - \frac{r_2(r_1 - r_2)}{\lambda(r_1 - r_2) + r_1 r_2 \mu} e^{-\mu K}\Bigg]^{-1}. \end{aligned} \tag{21}$$

The expressions for Q_1 and Q_2 follow directly from (20) and (19).
Summarizing, we have found that, in the M/M/1 queue with a two-step service speed function, the density of the "sojourn time" is given by (14) and (18), the constants Q_1, Q_2, Q_3

being specified by (19), (20) and (21). Observing that $S_n = W_n + B_n$, where W_n and B_n are independent, now yields the distribution of W, and hence, using PASTA, the steady-state workload distribution. We give $\mathbb{P}(W = 0)$ and the density $f_W(x),\ x > 0$:

$$\mathbb{P}(W = 0) = \frac{Q_1 + Q_2}{\mu}, \tag{22}$$

$$f_W(x) = \begin{cases} Q_1\rho_1 \mathrm{e}^{(\frac{\lambda}{r_1}-\mu)x} + Q_2(1+\rho_2)\mathrm{e}^{\frac{\lambda}{r_2}x}, & \text{for } 0 < x \leq K, \\ Q_3\rho_2 \mathrm{e}^{(\frac{\lambda}{r_2}-\mu)x}, & \text{for } x > K. \end{cases} \tag{23}$$

Remark 3.1. *Note that the equations reduce to familiar results for the M/M/1 queue with service speed r_2 in case either $K = 0$, or $r_1 = r_2$. In particular, we then have*

$$f_S(x) = \mu(1-\rho_2)e^{-\mu(1-\rho_2)x}, \qquad \text{for } \ x > 0.$$

4 General service requirements

In this section we apply the procedure described in Section 2 to the general M/G/1 queue. The basic ideas are similar as in the M/M/1 case of Section 3. Again, we start the algorithm with Equation (11), which is the result of Step 1 in Section 2.

Step 2: *Sojourn time distribution on* $(0, K]$
The transforms in this step can be treated in a similar manner as the transforms in the exponential case of Section 3. First, to describe the inverse of $F_1(\omega)$, we define

$$H(x) := \beta^{-1} \int_0^x (1 - B(y))\mathrm{d}y$$

as the stationary residual service requirement distribution. Similar to [6, 7], let $\delta_1 = 0$ for $\rho_1 \leq 1$ and for $\rho_1 > 1$ let δ_1 be the unique positive zero of the function

$$\int_0^\infty e^{-xy}\rho_1 \mathrm{d}H(y) - 1.$$

Then, for $x > 0$, define

$$L(x) := \int_0^x e^{-\delta_1 y}\rho_1 \mathrm{d}H(y),$$

and

$$W_1(x) := \int_{0^-}^x e^{\delta_1 y}\mathrm{d}\left\{\sum_{n=0}^\infty L^{n^*}(y)\right\},$$

where $L^{n^*}(\cdot)$ denotes the n-fold convolution of $L(\cdot)$ with itself. Finally, let

$$S_1(x) := (1-\rho_1)\int_0^x B(x-y)\mathrm{d}W_1(y),$$

be the convolution of $(1-\rho_1)W_1(\cdot)$ with $B(\cdot)$. It may be checked that, as in [6, 7], the LST of $S_1(\cdot)$ equals $F_1(\omega)$, that is, Equation (3) with $i = 1$.
For $\rho_1 < 1$, we note that $(1-\rho_1)W_1(\cdot)$ and $S_1(\cdot)$ are the steady-state waiting-time and sojourn-time distributions in an M/G/1 queue with service speed r_1. In case $\rho_1 \geq 1$, $W_1(\cdot)$

may be interpreted in terms of a dam with release rate r_1 and capacity K. Specifically, the stationary waiting-time distribution for such a dam equals $W_1(\cdot)/W_1(K)$, see for instance [6], or [8], p. 536.
To obtain the sojourn time distribution on $(0, K]$, we apply Laplace inversion to each of the transforms in (11) as in Section 3. The inverse of the first LST $F_1(\omega)$ is described above. For the second transform

$$F_1(\omega)\mathbb{E}\left[e^{-\omega(S-r_2A)}I(S-r_2A>K)\right],$$

we recall that this involves a product of two LSTs, corresponding to the sum of a random variable with mass on $[0,\infty)$, and one with mass on $[K,\infty)$. Thus the sum has no mass on $(0,K]$. Using (13) for the third transform in (11) as in Section 3, we obtain, for $\rho_1 \neq 1$,

$$\mathbb{P}(S<x)=\frac{W(0)}{1-\rho_1}S_1(x)+\frac{(1-\frac{r_2}{r_1})}{1-\rho_1}\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S<y)\int_0^x S_1(x-y)f(y)\mathrm{d}y. \tag{24}$$

The above equation may be rewritten into an intuitively more appealing expression by using the interpretation of $f(\cdot)$. As discussed in Step 1, the event $S_n \leq K$ implies that either the previous sojourn time was also at or below K, or a downcrossing has occurred between the two consecutive arrivals. Denote the probability of a downcrossing of K between two successive arrivals by $P_{\downarrow K}$. Then, obviously,

$$P_{\downarrow K}=\int_K^\infty e^{-\frac{\lambda}{r_2}(y-K)}\mathrm{d}\mathbb{P}(S<y).$$

Let A_λ be a generic exponential random variable with mean $1/\lambda$. It is then easily seen that

$$\mathbb{E}\left[e^{-\omega(K-A_{\lambda/r_2})^+}\right]=\frac{\lambda}{r_2\omega-\lambda}\left(e^{-\frac{\lambda}{r_2}K}-e^{-\omega K}\right)+e^{-\frac{\lambda}{r_2}K}.$$

In case $\rho_1 < 1$, let $\hat{S}_1$ denote a generic sojourn time in an M/G/1 queue with service rate r_1. Combining the above directly gives, for $x \in (0, K]$ and $\rho_1 < 1$,

$$\mathbb{P}(S<x)=\frac{Q}{1-\rho_1}\mathbb{P}(\hat{S}_1<x)+\frac{1-\frac{r_2}{r_1}}{1-\rho_1}P_{\downarrow K}\mathbb{P}(\hat{S}_1+(K-A_{\lambda/r_2})^+<x), \tag{25}$$

where

$$Q:=\int_0^K e^{-\frac{\lambda}{r_1}y}\mathrm{d}\mathbb{P}(S<y)+\frac{r_2}{r_1}\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S<y).$$

To provide some insight, let a cycle be the sample path in $(0, K]$ starting when the workload process enters $(0, K]$ and ending when it leaves $(0, K]$. Then, the two probabilities in (25) have a direct interpretation: The first probability stems from sojourn times of customers arriving in cycles starting from the empty system, while the second term is due to cycles starting with a downcrossing of K. The sum with $(K - A_{\lambda/r_2})^+$ in the second probability corresponds to the first "waiting time" after such a downcrossing.
Finally, in case $\rho_1 \geq 1$ the intuitive form may be expressed in a similar way as (25). In that case, let $\hat{W}_1$ be a generic waiting time in an M/G/1 dam with service speed r_1 and finite buffer K and let B be a generic service requirement. Expression (25) then holds upon replacing $\hat{S}_1$ by $\hat{W}_1 + B$ and $1/(1 - \rho_1)$ by $W_1(K)$.

Step 3: *Sojourn time distribution on* (K,∞)

Taking the LST of (24) on $(0,K]$ and substituting the result in (4) yields $\phi(\omega)$. Below, we apply Equation (4) directly though to derive the sojourn time distribution on (K,∞). First, define

$$W_2(x) := (1-\rho_2)\sum_{n=0}^{\infty}\rho_2^n H^{n^*}(x).$$

Because $\rho_2 < 1$, $W_2(\cdot)$ corresponds to the steady-state waiting-time distribution in an M/G/1 queue with service speed r_2. Let $S_2(x) = W_2(x) * B(x)$ be the stationary sojourn time distribution in such a queue, with generic random variable $\hat{S}_2$. As is well-known, $F_2(\omega)$ in (3) is the LST of $S_2(\cdot)$.

For convenience, denote $\gamma(\omega) := \int_0^K e^{-\omega x}\mathrm{d}\mathbb{P}(S<x)$. Using standard algebra, we deduce

$$\lambda\frac{\gamma(\lambda)-\gamma(\omega)}{\omega-\lambda} = \mathbb{E}\left[e^{-\omega(S-A_\lambda)^+}I(S\le K)\right] - \gamma(\lambda). \tag{26}$$

Define, for $0 \le x \le K$,

$$\begin{aligned}\tilde{S}(x) &:= \mathbb{P}((S-A_{\lambda/r_1})^+ I(S\le K)\le x)\\ &= \int_0^x \tilde{s}(y)\mathrm{d}y + \tilde{S}(0),\end{aligned}$$

where $\tilde{S}(0) = \int_0^K e^{-\frac{\lambda}{r_1}y}\mathrm{d}\mathbb{P}(S<y)$, which is also equal to $\gamma(\lambda/r_1)$, and

$$\tilde{s}(x) := \int_x^K \frac{\lambda}{r_1}e^{-\frac{\lambda}{r_1}(y-x)}\mathrm{d}\mathbb{P}(S<y).$$

Combining the above with (4) rewritten as

$$\phi(\omega) = F_2(\omega)\frac{W(0)}{1-\rho_2} + F_2(\omega)\frac{1-\frac{r_1}{r_2}}{1-\rho_2}\frac{\lambda/r_1}{\omega-\lambda/r_1}\left(\gamma\left(\lambda/r_1\right)-\gamma(\omega)\right),$$

we obtain, for $x > K$,

$$\mathbb{P}(S<x) = \frac{W(0)}{1-\rho_2}S_2(x) + \frac{1-\frac{r_1}{r_2}}{1-\rho_2}\int_0^K S_2(x-y)\tilde{s}(y)\mathrm{d}y. \tag{27}$$

Alternatively, using that

$$W(0) = \frac{r_1}{r_2}Q + (1-\frac{r_1}{r_2})\gamma(\lambda/r_1),$$

the sojourn time distribution may be expressed as

$$\mathbb{P}(S<x) = \frac{\frac{r_1}{r_2}Q}{1-\rho_2}\mathbb{P}(\hat{S}_2<x) + \frac{1-\frac{r_1}{r_2}}{1-\rho_2}\mathbb{P}(\hat{S}_2 + (S-A_{\lambda/r_1})^+ I(S\le K) < x). \tag{28}$$

Here, the first probability relates to busy cycles in which all "sojourn times" are larger than K. In that case, the system is identical to an M/G/1 queue with service speed r_2. In case $S_n \le K$ before the end of the busy cycle, the sample path above level K in the subsequent part of the busy cycle is initiated by $S - A_{\lambda/r_1}$ with $S \le K$, as is reflected in the second term. Note that Equation (2.15) in [7] has a similar structure.

Step 4: *Determination of the constants*
Using the fact that $\lim_{x\to\infty}\mathbb{P}(S<x)=1$ and $\lim_{x\to\infty}S_2(x)=1$, we deduce from (27) that

$$W(0)=1-\rho_2-(1-\frac{r_1}{r_2})\left(\mathbb{P}(S<K)-\int_0^K e^{-\frac{\lambda}{r_1}y}\mathrm{d}\mathbb{P}(S<y)\right). \qquad (29)$$

Moreover, substituting $x=K$ in (24) yields

$$\mathbb{P}(S<K)=\frac{W(0)}{1-\rho_1}S_1(K)+\frac{(1-\frac{r_2}{r_1})}{1-\rho_1}\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S<y)\int_0^K S_1(K-y)f(y)\mathrm{d}y. \qquad (30)$$

Equations (10) and (24) can be used to determine the constants $\int_0^K e^{-\frac{\lambda}{r_1}y}\mathrm{d}\mathbb{P}(S<y)$ and $\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S<y)$ in terms of $W(0)$ and $\mathbb{P}(S<K)$. Hence, using (29) and (30), we find after lengthy calculations that

$$W(0) = \frac{(1-\rho_1)(1-\rho_2)(D_1+e^{-\frac{\lambda}{r_1}K}f_2)}{(1-\frac{r_1}{r_2})S_1(K)D_2+D_3+(1-\rho_1)\frac{r_1}{r_2}e^{-\frac{\lambda}{r_1}K}f_2}, \qquad (31)$$

$$\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S<y) = W(0)\frac{D_1-e^{-\frac{\lambda}{r_1}K}S_1(K)}{D_1+e^{-\frac{\lambda}{r_1}K}f_2}, \qquad (32)$$

where

$$\begin{aligned}
f_i &:= \int_0^K \frac{\lambda}{r_i}e^{\frac{\lambda}{r_i}(K-y)}S_1(y)\mathrm{d}y, \qquad i=1,2,\\
D_1 &:= 1-\rho_1-e^{-\frac{\lambda}{r_1}K}f_1,\\
D_2 &:= D_1+e^{-\frac{\lambda}{r_1}K}\left(\frac{r_2}{r_1}f_2-(1-\rho_1)\right),\\
D_3 &:= D_1\left(1-\rho_1+(1-\frac{r_1}{r_2})(1-\frac{r_2}{r_1})f_2\right).
\end{aligned}$$

Summarizing, the density of the "sojourn time" is given by (24) and (27) (see (25) and (28) for another representation), where the main constants are given by (31) and (32). Because $S_n=W_n+B_n$, where W_n and B_n are independent, we also directly obtain the "waiting-time" distribution and, applying PASTA, the steady-state workload distribution. In particular, for $x\in(0,K]$, we have

$$\mathbb{P}(W<x)=W(0)W_1(x)+(1-\frac{r_2}{r_1})\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S<y)\int_0^x W_1(x-y)f(y)\mathrm{d}y, \qquad (33)$$

and for $x>K$,

$$\mathbb{P}(W<x)=\frac{W(0)}{1-\rho_2}W_2(x)+\frac{1-\frac{r_1}{r_2}}{1-\rho_2}\int_0^K W_2(x-y)\tilde{s}(y)\mathrm{d}y.$$

Note that we may determine the density $\tilde{s}(y)$, $0<y\le K$, up to some constants, once we have found the workload distribution on $(0,K]$.

5 Special cases and extensions

In this section we first consider some special cases of the model with a two-step service rule and conclude with the extension to the N-step service speed function. The case of exponentially distributed service requirements and the two-step service rule has already been treated in Section 3. In Subsection 5.1 we focus on service requirements with a rational LST to provide some structural properties. Furthermore, by allowing general service requirements, but letting $r_2 \to \infty$ we obtain an M/G/1 queue with disasters (clearings) at level crossings in Subsection 5.2. Finally, in Subsection 5.3 we analyze the M/G/1 queue with an N-step service speed function.

5.1 Service requirements with rational LST

In this subsection we assume that the LST $\beta(\omega)$ is a rational function of ω. This allows us to obtain some structural properties of the steady-state sojourn time distribution. In particular, let

$$\beta(\omega) = \frac{\beta_1(\omega)}{\beta_2(\omega)},$$

where $\beta_1(\omega)$ and $\beta_2(\omega)$ are polynomials in ω with $\beta_2(\omega)$ of degree n and $\beta_1(\omega)$ of degree strictly less than n (in other words, we assume $B(0^+) = 0$). This class includes, for instance, phase-type distributions. We use the notation M/K_n/1 to denote single-server queues where the service requirements have such rational LSTs.

The inverse of $F_i(\omega)$, $i = 1, 2$, can now be given more explicitly. Rewrite (3) as

$$F_i(\omega) = (1 - \rho_i)\frac{r_i\beta_1(\omega)}{r_i\beta_2(\omega) - \lambda(\beta_2(\omega) - \beta_1(\omega))/\omega}.$$

Let $\delta_2 := 0$ and $\epsilon > 0$ be arbitrary small. It then follows from Rouché's theorem applied to the function $r_i\beta_2(\omega) - \lambda(\beta_2(\omega) - \beta_1(\omega))/\omega$ for Re $\omega \leq \delta_i + \epsilon$, $i = 1, 2$, that the function has exactly n zeros in the plane with Re $\omega < \delta_i + \epsilon$ (see for instance [8], p. 323, in case $\rho_i < 1$).

For ease of presentation, we assume that the function $r_i\beta_2(\omega) - \lambda(\beta_2(\omega) - \beta_1(\omega))/\omega$, $i = 1, 2$, has one zero of multiplicity m_i, $m_i = 2, 3, \dots, n$, while the other $n - m_i$ zeros are simple, i.e., have multiplicity one. Let $\omega_i(1)$ be the non-simple zero and $\omega_i(m_i + 1), \dots, \omega_i(n)$ be the distinct simple zeros. By a partial-fraction expansion and Laplace inversion of $F_i(\omega)$, we have

$$s_i(x) = \sum_{k=1}^{m_i} \tilde{Q}_i(k)x^k e^{\omega_i(1)x} + \sum_{k=m_i+1}^{n} \tilde{Q}_i(k)e^{\omega_i(k)x},$$

for some constants $\tilde{Q}_i(k)$, $i = 1, 2$ and $k = 1, \dots, n$. In other words, the density of the sojourn time in the M/K_n/1 queue with service speed r_i may be written as the mixture of m_i Erlang densities with scale parameter $\omega_i(1)$ and $n - m_i$ exponential terms.

It now follows from the general expressions in Section 4 that the "sojourn time" density has a similar structure. First consider $0 < x \leq K$. Note that the convolution of an Erlang(k, μ) distribution with an exponential term is a mixture of Erlang(i, μ), $i = 1, \dots, k$, distributions and the same exponential. Using (24), we obtain, for $0 < x \leq K$,

$$f_S(x) = \sum_{k=1}^{m_1} Q_1(k)x^k e^{\omega_1(1)x} + \sum_{k=m_1+1}^{n} Q_1(k)e^{\omega_1(k)x} + Q_0 e^{\frac{\lambda}{r_2}x}.$$

Observe that $f_S(x)$ has the same Erlang and exponential terms as the sojourn time density in an ordinary M/K_n/1 queue with service speed r_1 (for $\rho_1 < 1$) plus one additional exponential $\exp(x\lambda/r_2)$ (but with different constants). Further observe that $\omega_i(k)$, $i = 1, 2$, $k = m_i + 1, \ldots, n$, might be complex, in which case its complex conjugate will also appear, leading to an exponential times a cosine, respectively, sine function.
Second, for $x > K$, we use the fact that the conditional sojourn time density of $\hat{S}_2$ has the same structure as the density of $\hat{S}_2$ itself, i.e.,

$$s_2(x + y|\hat{S}_2 > y) = \sum_{k=1}^{m_2} \hat{Q}_2(k)x^k e^{\omega_2(1)x} + \sum_{k=m_2+1}^{n} \hat{Q}_2(k)e^{\omega_2(k)x},$$

for some constants $\hat{Q}_2(k)$, $k = 1, \ldots, n$ (which depend on y). Combining the above with (27), we deduce that

$$f_S(x) = \sum_{k=1}^{m_2} Q_2(k)x^k e^{\omega_2(1)x} + \sum_{k=m_2+1}^{n} Q_2(k)e^{\omega_2(k)x}.$$

Finally, using the normalization condition $\int_0^\infty f_S(x)\mathrm{d}x = 1$ together with the definitions of $Q_i(k)$, $i = 1, 2$ and $k = 1, \ldots, n$, provides $2n + 1$ equations for determining the $2n + 1$ constants Q_0, $Q_i(k)$, for $i = 1, 2$ and $k = 1, \ldots, n$.

5.2 Disasters at level crossings

A special case of the model discussed in Section 4 is an M/G/1 queue with disasters at level crossings, see e.g. [5]. In such a model, the system is immediately cleared when the workload exceeds some level K, that is, the residual amount of work is removed from the system when the workload becomes larger than K. In case $r_2 \to \infty$ in our model, the available amount of work is not removed but served instantaneously when the workload upcrosses K. However, both interpretations of the work present after such an upcrossing result in identical mathematical models.
First, we note that the workload embedded at epochs right after arrival instants may be larger than K in our model (with $r_2 \to \infty$). In terms of clearing processes, this embedded workload may be considered as the overshoot (and thus the amount of work lost) rather than the actual amount of work present. Letting $r_2 \to \infty$ in (27) yields, for $x > K$,

$$\mathbb{P}(S < x) = W(0)B(x) + \int_0^K B(x - y)\tilde{s}(y)\mathrm{d}y,$$

where $\tilde{s}(\cdot)$ may, for instance, be determined by letting $r_2 \to \infty$ in (24).
For clearing models, the workload might be a more natural performance measure than the "sojourn time". In particular, we have $\mathbb{P}(W \leq K) = 1$ and, for $x \in (0, K)$, Equation (33) reduces to

$$\mathbb{P}(W < x) = W(0)W_1(x) - \mathbb{P}(S > K)\frac{\lambda}{r_1}\int_0^x W_1(y)\mathrm{d}y.$$

By letting $r_2 \to \infty$ in (31) and (32), we obtain the two main constants

$$W(0) = \frac{(1-\rho_1)D_1}{S_1(K)\left(D_1 - e^{-\frac{\lambda}{r_1}K}D_4\right) + D_1 D_4},$$

$$\mathbb{P}(S > K) = W(0)\frac{D_1 - e^{-\frac{\lambda}{r_1}K}S_1(K)}{D_1},$$

where

$$D_4 = 1 - \rho_1 - \frac{\lambda}{r_1}\int_0^K S_1(y)\mathrm{d}y.$$

Observe that Equations (10) and (29) are identical when $r_2 \to \infty$. Because $\mathbb{P}(S > K)$ equals $\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S < y)$ in that case, the three constants can also be found from the three independent equations as discussed in Section 4.

Remark 5.1. *In the M/M/1 case with $r_1 = 1$, it may be checked that (22) and (23) for $r_2 \to \infty$, or the expressions given above, indeed reduce to the workload density and the probability of an empty system of [5, Theorem 3].*

5.3 N-step service rule

In this subsection we extend the analysis to an N-step service rule. Specifically, let $r(x) = r_i$ for $x \in (K_{i-1}, K_i]$, $i = 1, \dots, N$ (where $K_0 = 0$ and $K_N = \infty$). Also, define $\rho_i := \lambda\beta/r_i$. For stability, we require that $\rho_N < 1$. The basic ideas are now similar to the case $N = 2$ discussed in Section 4.

Below, we give the derivation of the "sojourn time" distribution for the N-step service rule along similar lines as the four-step procedure described in Section 2. That is, we first present N different equations for $\phi(\omega)$. Second, we use a similar interpretation as in Step 1 to rewrite the N equations. Third, similar to Step 2 in Section 4 we analyze $\mathbb{P}(S < x)$ for $x \in (0, K_1]$. Then, we recursively determine $\mathbb{P}(S < x)$ for $x \in (K_{i-1}, K_i]$, $i = 2, \dots, N$ (comparable with Step 3). We conclude with some remarks about the determination of the constants.

Concerning the equations for $\phi(\omega)$, it follows from (1), (6), and conditioning on S_n that

$$\begin{aligned}
\mathbb{E}\left[e^{-\omega S_{n+1}}\right] &= \int_0^\infty \mathbb{E}\left[e^{-\omega S_{n+1}} | S_n = x\right]\mathrm{d}\mathbb{P}(S_n < x) \\
&= \beta(\omega)\sum_{j=1}^N \Bigg[\frac{\omega r_j}{\omega r_j - \lambda}\int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j}x}\mathrm{d}\mathbb{P}(S_n < x) \\
&\qquad - \frac{\lambda}{\omega r_j - \lambda}\int_{K_{j-1}}^{K_j} e^{-\omega x}\mathrm{d}\mathbb{P}(S_n < x)\Bigg],
\end{aligned}$$

with obvious modification for $\omega = \lambda/r_j$, $j = 1, \dots, N$. Using similar manipulations as in the proof of Lemma 2.1, we obtain N alternative equations for $\phi(\omega)$; for $i = 1, \dots, N$, we have

$$\begin{aligned}\phi(\omega) &= F_i(\omega)\frac{W(0)}{1-\rho_i} \qquad (34)\\ &+ \frac{F_i(\omega)}{1-\rho_i}\sum_{j=i+1}^{N}\left[\frac{\lambda(1-\frac{r_j}{r_i})}{\omega r_j-\lambda}\left(\int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j}x}\mathrm{d}\mathbb{P}(S<x) - \int_{K_{j-1}}^{K_j} e^{-\omega x}\mathrm{d}\mathbb{P}(S<x)\right)\right]\\ &+ \frac{F_i(\omega)}{1-\rho_i}\sum_{j=1}^{i-1}\left[\frac{\lambda(1-\frac{r_j}{r_i})}{\omega r_j-\lambda}\left(\int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j}x}\mathrm{d}\mathbb{P}(S<x) - \int_{K_{j-1}}^{K_j} e^{-\omega x}\mathrm{d}\mathbb{P}(S<x)\right)\right],\end{aligned}$$

with obvious notation for $F_i(\omega)$ and

$$W(0) = \sum_{j=1}^{N}\int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j}x}\mathrm{d}\mathbb{P}(S<x). \qquad (35)$$

In the remainder, we follow the convention that empty sums are equal to zero.

Step 1: *Rewriting (34)*
Fix some $i = 1,\dots,N$ and consider the term on the second line of (34). As in Step 1 of Section 2, $S_n > K_i$ and $S_{n+1} \leq K_i$ means that a downcrossing of level K_i occurs between the arrival epochs of customers n and $n+1$. Again, the residual interarrival time at a downcrossing of K_i is still exponential, but the service speed now depends on the value of S_n. In particular, the precise distribution of S_n on (K_i,∞) does not directly affect the distribution of $S_{n+1} \leq K_i$ but determines the service speed until the next arrival epoch. Using similar calculations as in Step 1 of Section 2, we obtain

$$\begin{aligned}&\sum_{j=i+1}^{N}\left[\frac{\lambda}{\omega r_j-\lambda}\left(\int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j}x}\mathrm{d}\mathbb{P}(S_n<x) - \int_{K_{j-1}}^{K_j} e^{-\omega x}\mathrm{d}\mathbb{P}(S_n<x)\right)\right]\\ &= \mathbb{E}\left[e^{-\omega(S_n-r(S_n)A_n)^+}I(S_n>K_i)\right] - \sum_{j=i+1}^{N}\int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j}x}\mathrm{d}\mathbb{P}(S_n<x)\\ &= \mathbb{E}\left[e^{-\omega(S_n-r(S_n)A_n)}I(S_n-r(S_n)A_n>K_i)\right]\\ &\quad + \sum_{j=i+1}^{N}\frac{\lambda}{\omega r_j-\lambda}\left(1-e^{-\omega K_i+\frac{\lambda}{r_j}K_i}\right)\int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j}x}\mathrm{d}\mathbb{P}(S_n<x).\end{aligned}$$

For convenience, we define the quantity

$$C_j := \int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j}x}\mathrm{d}\mathbb{P}(S<x),$$

which is clearly independent of ω. Then, by letting $n \to \infty$, we may rewrite (34) as

$$
\begin{aligned}
\phi(\omega) &= F_i(\omega)\frac{W(0)}{1-\rho_i} \qquad (36)\\
&+ \frac{F_i(\omega)}{1-\rho_i}\mathbb{E}\left[e^{-\omega(S-r(S)A)}I(S-r(S)A>K_i)\right]\\
&+ \frac{F_i(\omega)}{1-\rho_i}\sum_{j=i+1}^{N}(1-\frac{r_j}{r_i})C_j\frac{\lambda}{\omega r_j-\lambda}\left(1-e^{-\omega K_i+\frac{\lambda}{r_j}K_i}\right),\\
&+ \frac{F_i(\omega)}{1-\rho_i}\sum_{j=1}^{i-1}\left[\frac{\lambda(1-\frac{r_j}{r_i})}{\omega r_j-\lambda}\left(\int_{K_{j-1}}^{K_j}e^{-\frac{\lambda}{r_j}x}\mathrm{d}\mathbb{P}(S<x)-\int_{K_{j-1}}^{K_j}e^{-\omega x}\mathrm{d}\mathbb{P}(S<x)\right)\right]\\
&=: I+II+III+IV.
\end{aligned}
$$

Note that the intuitive observations made above are reflected in Terms II and III.

Step 2: *Sojourn time distribution on* $(0, K_1]$
First we consider $i = 1$, i.e., the interval $(0, K_1]$. Note that this implies that $IV = 0$.
As in Step 2 of Section 4 we now apply Laplace inversion to each of the Terms I, II, and III separately. Again, $S_1(\cdot)W(0)/(1-\rho_1)$ is the inverse of Term I, see also Section 4. Term II involves the convolution of two random variables, one with mass on $[0,\infty)$ and one with mass on (K_1,∞). Hence, the sum clearly has no mass on $(0, K_1]$.
For Term III, we note that $\frac{\lambda}{\omega r_j-\lambda}\left(1-e^{-\omega K_i+\frac{\lambda}{r_j}K_i}\right)$ is the Laplace Transform of the function

$$
f_{i,j}(x) = \begin{cases} \frac{\lambda}{r_j}e^{\frac{\lambda}{r_j}x}, & \text{for } 0 < x \le K_i,\\ 0, & \text{for } x > K_i. \end{cases}
$$

To provide some intuition, suppose that $S_n \in (K_{j-1}, K_j]$ and a downcrossing of level $K_i \le K_{j-1}$ occurs in the subsequent interarrival time, which has stationary probability

$$
P^j_{\downarrow K_i} = \int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j}(y-K_i)}\mathrm{d}\mathbb{P}(S<y).
$$

Then $P^j_{\downarrow K_i}f_{i,j}$ may be interpreted as C_j times the "density" of $(K_i - A_{\lambda/r_j})^+$ (in fact, $(K_i - A_{\lambda/r_j})^+$ has a defective distribution with an atom in 0).
Combining the above and applying Laplace inversion provides an extension of Equation (24) to the case of an N-step service rule, with $0 < x \le K_1$,

$$
\mathbb{P}(S<x) = \frac{W(0)}{1-\rho_1}S_1(x) + \frac{1}{1-\rho_1}\sum_{j=2}^{N}(1-\frac{r_j}{r_1})C_j\int_{0^+}^{x}S_1(x-y)f_{1,j}(y)\mathrm{d}y. \qquad (37)
$$

Note that the difference with $N = 2$ is the fact that the service speed now depends on the previous "sojourn time" in case of a downcrossing of K_1. This naturally leads to a mixture of convolutions of $S_1(\cdot)$ with various exponential functions depending on the service speed in the second part of (37).

Step 3: *Sojourn time distribution on* $(K_{i-1}, K_i]$
In Step 2 we obtained the "sojourn time" distribution on the first interval $(0, K_1]$. We

may now recursively determine the "sojourn time" distribution on the remaining intervals. That is, suppose that $\mathbb{P}(S < x)$ is known for $x \in (K_{j-1}, K_j]$, $j = 1, \ldots, i-1$, with $i = 2, \ldots, N$ (the case $i = 1$ corresponds to Step 2). Using (36), we then find $\mathbb{P}(S < x)$ for $x \in (K_{i-1}, K_i]$.

To do so, we apply Laplace inversion again to each of the four terms in (36). Terms I, II, and III can be treated as in Step 2, with obvious notation for $W_i(\cdot)$, $i = 2, \ldots, N$. For the fourth term, we apply similar arguments as in Step 3 of Section 4, in particular Equation (26). Thus,

$$IV = \frac{F_i(\omega)}{1-\rho_i} \sum_{j=1}^{i-1} (1 - \frac{r_j}{r_i}) \left(\mathbb{E}\left[e^{-\omega(S - A_{\lambda/r_j})^+} I(K_{j-1} < S \leq K_j) \right] - C_j \right).$$

Note again that $(S - A_{\lambda/r_j})^+ I(K_{j-1} < S \leq K_j)$ has a defective distribution function with an atom at zero, $\tilde{S}_j(0) := C_j$. Moreover, the density reads, for $0 < x < K_j$,

$$\tilde{s}_j(x) := \int_{\max(x, K_{j-1})}^{K_j} \frac{\lambda}{r_j} e^{-\frac{\lambda}{r_j}(y-x)} \mathrm{d}\mathbb{P}(S < y).$$

Because we assumed that $\mathbb{P}(S < x)$ is known on $(0, K_{i-1}]$, $\tilde{s}_j(x)$ is computable for every $j = 1, \ldots, i-1$.

Now, combining the above and applying Laplace inversion to (36) yields, for $K_{i-1} < x \leq K_i$, $i = 1, \ldots, N$,

$$\begin{aligned} \mathbb{P}(S < x) \quad &= \quad \frac{W(0)}{1-\rho_i} S_i(x) + \frac{1}{1-\rho_i} \sum_{j=i+1}^{N} (1 - \frac{r_j}{r_i}) C_j \int_{0^+}^{x} S_i(x-y) f_{i,j}(y) \mathrm{d}y \\ &\quad + \frac{1}{1-\rho_i} \sum_{j=1}^{i-1} (1 - \frac{r_j}{r_i}) \int_{0^+}^{K_j} S_i(x-y) \tilde{s}_j(y) \mathrm{d}y. \end{aligned} \tag{38}$$

The $S_i(\cdot)$ term and the convolution of $S_i(\cdot)$ with $\tilde{s}_j(\cdot)$ are similar to the case $N = 2$, see (27). For $i = 1, \ldots, N-1$, we just have an additional convolution of $S_i(\cdot)$ with $f_{i,j}(\cdot)$, which is the consequence of "sojourn times" after a downcrossing of K_i, as discussed in Step 2.

Step 4: *Determination of the constants*
Taking $i = N$ and letting $x \to \infty$ in (38), yields

$$W(0) = 1 - \rho_N - \sum_{j=1}^{N-1} (1 - \frac{r_j}{r_N}) \left(\mathbb{P}(K_{j-1} \leq S < K_j) - C_j \right). \tag{39}$$

Moreover, (38) can be used to give expressions for $\mathbb{P}(S < K_i)$ and C_i, $i = 1, \ldots, N-1$. To obtain the latter $N-1$ constants, differentiate (38) with respect to x, multiply by $\exp(-\lambda x / r_i)$, and integrate over the interval $(K_{i-1}, K_i]$. Together with (35) and (39), this provides $2N$ independent equations to determine the $2N$ unknowns: $W(0)$, $\mathbb{P}(S < K_i)$ for $i = 1, \ldots, N-1$, and C_i, $i = 1, \ldots, N$.

Acknowledgement

The authors are indebted to Johan van Leeuwaarden for interesting discussions and to Prof. Ton de Kok for posing the problem.
A preliminary version of this paper was presented in a Korea-Netherlands workshop in Seoul [4]; the second author greatfully acknowledges the hospitality of Prof. B.D. Choi.

References

[1] Asmussen, S. (2003). *Applied Probability and Queues*, Second Edition. Springer, New York.

[2] Bae, J., S. Kim, E.Y. Lee (2003). Average cost under the $P^M_{\lambda,\tau}$ policy in a finite dam with compound Poisson inputs. *Journal of Applied Probability* **40**, 519–526.

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

[4] Bekker, R., and O.J. Boxma (2005). *Queues with adaptable service speed*. In: B.D. Choi (ed.), Korea-Netherlands; Joint conference on Queueing theory and its Applications to Telecommunication Systems, 91–100.

[5] Boxma, O.J., D. Perry, W. Stadje (2001). Clearing models for M/G/1 queues. *Queueing Systems* **38**, 287–306.

[6] Cohen, J.W. (1976). On Regenerative Processes in Queueing Theory. *Lecture Notes in Economics and Mathematical Systems* **121**. Springer-Verlag, Berlin.

[7] Cohen, J.W. (1976). On the optimal switching level for an M/G/1 queueing system. *Stochastic Processes and Their Applications* **4**, 297–316.

[8] Cohen, J.W. (1982). *The Single Server Queue*, North-Holland, Amsterdam.

[9] Feinberg, E.A., and O. Kella (2002). Optimality of D-policies for an M/G/1 queue with a removable server. *Queueing Systems* **42**, 355–376.

[10] Gaver, D.P., and R.G. Miller (1962). Limiting distributions for some storage problems. In: *Studies in Applied Probability and Management Science*, 110–126.

[11] Harrison, J.M., and S.I. Resnick (1976). The stationary distribution and first exit probabilities of a storage process with general release rule. *Mathematics of Operations Research* **1**, 347–358.

[12] Lee, J., and J. Kim (2005). A workload-dependent M/G/1 queue under a two-stage service policy. *Operations Research Letters*, to appear.

[13] Moran, P.A.P. (1969). A theory of dams with continuous input and a general release rule. *Journal of Applied Probability* **6**, 88–98.

[14] Tijms, H.C. (1976). Optimal control of the workload in an M/G/1 queueing system with removable server. *Math. Operationsforsch. Statist.* **7**, 933–944.

[15] Tijms, H.C., and F.A. van der Duyn Schouten (1978). Inventory control with two switch-over levels for a class of M/G/1 queueing systems with variable arrival and service rate. *Stochastic Processes and Their Applications* **6**, 213–222.