J. Appl. Prob. **50,** 612–631 (2013)
Printed in England
© *Applied Probability Trust* 2013

A DUALITY APPROACH TO QUEUES WITH SERVICE RESTRICTIONS AND STORAGE SYSTEMS WITH STATE-DEPENDENT RATES

D. PERRY,* *University of Haifa*
W. STADJE,** *University of Osnabrück*
S. ZACKS,*** *Binghampton University*

Abstract

Based on pathwise duality constructions, several new results on truncated queues and storage systems of the G/M/1 type are derived by transforming the workload (content) processes into certain 'dual' M/G/1-type processes. We consider queueing systems in which (a) any service requirement that would increase the total workload beyond the capacity is truncated so as to keep the associated sojourn time below a certain constant, or (b) new arrivals do not enter the system if they have to wait more than one time unit in line. For these systems, we derive the steady-state distributions of the workload and the numbers of customers present in the systems as well as the distributions of the lengths of busy and idle periods. Moreover, we use the duality approach to study finite capacity storage systems with general state-dependent outflow rates. Here our duality leads to a Markovian finite storage system with state-dependent jump sizes whose content level process can be analyzed using level crossing techniques. We also derive a connection between the steady-state densities of the non-Markovian continuous-time content level process of the G/M/1 finite storage system with state-dependent outflow rule and the corresponding embedded sequence of peak points (local maxima).

Keywords: Queue with service restrictions; storage system; state-dependent rate; M/G/1; G/M/1; steady state; duality; level crossing; peak point

2010 Mathematics Subject Classification: Primary 60K25
Secondary 90B22

1. Introduction

The purpose of this paper is to present a prototype duality approach to M/G-type and G/M-type queues and dams. We show how certain pathwise duality techniques can be used to analyze queueing systems with admission controls and storage systems with state-dependent outflow rates. Consider, for example, the workload or content process of a G/M/1 system with some restrictions at its capacity limit, which has renewal arrival times and exponentially distributed jumps. The duality transforms the sample paths of such a process into those of an M/G/1-type process. The paths are turned upside down, possibly modified on intervals of constancy, jumps are replaced by linear segments, and deterministic pieces between two

Received 27 May 2010; revision received 6 September 2012.

* Postal address: Department of Statistics, University of Haifa, 31905 Haifa, Israel.
Email address: dperry@stat.haifa.ac.il
** Postal address: Fachbereich Mathematik/Informatik, University of Osnabrück, 49069 Osnabrück, Germany.
Email address: wolfgang@mathematik.uni-osnabrueck.de
*** Postal address: Department of Mathematical Sciences, Binghampton University, Binghampton, NY 13902-6000, USA. Email address: shelly@math.binghampton.edu

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

jumps become jumps of the appropriate size. The resulting process again represents a system with different restrictions at the capacity limit and other features (such as no idle periods). However, it is of the M/G/1 type and, therefore, Markovian, so it can be studied by well-known methods. Due to the special form of the admission restrictions, every application of the duality requires special modifications. We present the method in two important examples, one from queueing and one from storage theory, which may serve as *prototypes* for the application of the duality to specific models. We believe that the approach introduced here will lay the groundwork for other stochastic storage models, insurance risk, and reliability systems.

We provide an overview of our queueing and then our storage results below.

1.1. Queueing

Many single-server systems have to deal with excessive service requirements in order to avoid overloads or to meet capacity constraints. The following two natural possibilities were suggested in [9], [23], and [25].

- *Queueing model 1.* Under the *truncated service policy*, any service requirement that would increase the total workload beyond some constant capacity threshold is reduced so that this threshold is reached but not exceeded. Note that, according to this policy, every arriving customer is admitted to the system.

- *Queueing model 2.* Under the *bounded waiting time policy*, new arrivals whose waiting times in line would exceed some fixed constant are not admitted to the system. According to this policy, admission is interrupted as long as the workload process stays above the threshold.

In the GI/G/1 case, queueing model 1 corresponds to the standard finite dam with constant release rule; various versions have been treated in [2], [3], [6], [10], [11], [12], [21], [24], and [27]. From the point of view of prospective customers, the two models also represent GI/G/1 with deterministic customer patience. In queueing model 1 each customer stays at most for a sojourn time of one time unit. In queueing model 2 the bound 1 is the maximum time a customer is willing to wait in line; this system is often called GI/G/1 + D, and we consider M/G/1 + D and G/M/1 + D.

For queueing model 2, the distribution for the workload in steady state and for the busy period length was derived in [19] and [20] for special cases with phase-type service times. In [16] a related inventory system was studied. For other variants of this model, see [6], [11], [13], and [26]. A busy period analysis of both models is given in [23] for the M/M/1 case and in [25] for the M/G/1 and G/M/1 case. Related questions were dealt with in [4] and [27].

The paper contains new results on the M/G/1 and the G/M/1 versions of both models with emphasis on the G/M/1 version. In Section 2 we consider queueing model 1 for M/G/1 and in particular derive the steady-state distributions of the sojourn time and the number of customers in the system. In Section 3 we study the M/G/1 version of queueing model 2, derive the steady-state distribution of the workload, and connect it to queueing model 1. In Section 4 we present the duality between the two systems, which allows us for the first time to study the G/M/1 version of queueing model 1. Through this approach we find the steady-state distributions of the workload and the number of customers as well as the distributions of the busy and idle periods and the times between overflows.

Alternative policies not considered in this paper are refusal of any customer whose *sojourn time* would be greater than some maximum value or quasirestricted accessibility (reducing the over-the-threshold parts of the service requirements by certain fractions). Results on the

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

steady-state distribution of the workload under these policies can be found in [5], [14], and [23] for M/G/1. The G/M/1 cases are open. The possible applicability of our method to these examples is discussed in the conclusion section.

1.2. Storage

In Section 5 we consider the duality for a storage system with finite capacity (a truncation threshold), renewal interarrival times for the inputs, and a general state-dependent release rate. For Poisson arrival times, this system was frequently studied; see [1] and the references given therein. Using the duality for the corresponding content process, we compute its steady-state distribution in the G/M/1 case. This approach leads to a *Markovian finite storage system with state-dependent jump sizes* whose content level process can be analyzed by level crossing techniques. In the last theorem of this section we consider the steady-state densities of the above G/M/1 content process and the corresponding sequence of its peak points (local maxima). Denote these densities by $f_{\boldsymbol{D}_{\mathrm{G/M/1}}}$ and $f_{\boldsymbol{PP}}$, respectively. It is proved that

$$r(x) f_{\boldsymbol{D}_{\mathrm{G/M/1}}}(x) = f_{\boldsymbol{PP}}(x) \tag{1}$$

for all positive x smaller than the capacity, where $r(x)$ is the outflow rate function. Equation (1) provides an interesting connection between the steady-state distribution of a non-Markovian continuous-time finite storage system (of G/M/1 type) and that of the subsequence of its peak points.

All steady-state distributions are given in closed form in terms of their densities in the form of a series of convolutions. In Section 6 we present an example to show how our duality methods can be applied to get more explicit formulae in special cases: we compute the Laplace–Stieltjes transform (LST) of the busy period of queueing model 1 in the case of Erlang arrivals. In the concluding remarks in Section 7 we summarize our approach and point to possible further applications.

2. Queueing model 1: basic facts for the M/G/1 case

Queueing model 1 is a single-server system of GI/G/1 type with a restriction on the sojourn time. Every customer is admitted to the system, but his/her service might be truncated to keep the sojourn time below a certain constant threshold, the maximum workload capacity, which we set equal to 1. Let $\nu \in (0, 1]$ be the initial workload at time $T_0 = 0$, let $0 < T_1 < T_2 < \cdots$ be the arrival times of customers, and assume that the interarrival times $T_n - T_{n-1}$, $n \in \mathbb{N}$, are independent and identically distributed (i.i.d.) random variables with distribution function H and mean $1/\lambda$, while the service requests $S_1, S_2, \dots$ at the times $T_1, T_2, \dots$ are independent of the T_i and i.i.d. with distribution function G and mean $1/\mu$. Each service request is known upon arrival. Under the truncated service admission rule, the workload process $\boldsymbol{V}^1 = \{V^1(t) \colon t \geq 0\}$ is defined step by step as

$$V^1(t) = \begin{cases} \nu, & t = 0, \\ \max[V^1(T_{n-1}) - (t - T_{n-1}), 0], & T_{n-1} \leq t < T_n,\ n \geq 1, \\ V^1(T_n-) + \bar{S}_n, & t = T_n,\ n \geq 1, \end{cases}$$

where $\bar{S}_n = \min[S_n, 1 - V^1(T_n-)]$ is the *actual service time* assigned to the nth customer. A piece of a typical sample path of $\boldsymbol{V}^1$ is depicted in Figure 1(a) in Section 4. It starts at time A_0 at full capacity $\nu = 1$ and has customer arrivals (positive jumps) at the time epochs $A_1, A_2, \dots$ with truncations at A_5 and A_6.

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

In the M/G/1 case (in which $H(t) = 1 - \mathrm{e}^{-\lambda t}$) the steady-state distribution function F_1 of $\boldsymbol{V}^1$ and the LST of the busy period distribution are known. The function F_1 has been derived by different methods [10], [12]; it is given by

$$F_1(x) = \frac{\sum_{n=0}^{\infty} \int_0^x \mathrm{e}^{\lambda(x-u)}[-\lambda(x-u)]^n \,\mathrm{d}G_n(u)/(n!)}{\sum_{n=0}^{\infty} \int_0^1 \mathrm{e}^{\lambda(1-u)}[-\lambda(1-u)]^n \,\mathrm{d}G_n(u)/(n!)}, \qquad 0 \le x \le 1, \tag{2}$$

where G_n, $n \ge 0$, denotes the n-fold convolution of the service time distribution G with itself. The busy period distribution is complicated; its LST was derived in [25].

Let us now determine some other important characteristics in the M/G/1 case, in particular the steady-state probability that there are n customers present, say p_n. We also need the steady-state distribution of the sojourn time. The following two propositions seem to be new.

Proposition 1. *Let Y_n^1 and W_n^1 be the sojourn time and the waiting time of the nth customer, respectively. Then* $\lim_{n\to\infty} \mathbb{P}(W_n^1 \le x) = F_1(x)$ *and*

$$\lim_{n\to\infty} \mathbb{P}(Y_n^1 \le x) = (G * F_1)(x), \qquad x < 1,$$
$$\lim_{n\to\infty} \mathbb{P}(Y_n^1 = 1) = 1 - (G * F_1)(1),$$

where '$*$' denotes the convolution operator.

Proof. By definition, $Y_n^1 = W_n^1 + \bar{S}_n$. Thus, for $x < 1$,

$$\lim_{n\to\infty} \mathbb{P}(Y_n^1 \le x) = \lim_{n\to\infty} \mathbb{P}(W_n^1 + \bar{S}_n \le x) = \lim_{n\to\infty} \mathbb{P}(W_n^1 + S_n \le x) = (G * F_1)(x),$$

since S_n is independent of W_n^1 and, by the PASTA property, F_1 is the steady-state distribution function of the waiting time. Finally, note that $Y_n^1 \le 1$.

Proposition 2. *Let $\boldsymbol{K} = \{K(t) : t \ge 0\}$ be the queue length process. Then*

$$p_n = \lim_{t\to\infty} \mathbb{P}(K(t) = n) = \int_0^1 \frac{\mathrm{e}^{-\lambda x}(\lambda x)^n}{n!} \,\mathrm{d}(G * F_1)(x) + \frac{\mathrm{e}^{-\lambda}\lambda^n}{n!}[1 - (G * F_1)(1)]. \tag{3}$$

Proof. The long-run average rate at which $\boldsymbol{K}$ makes a jump from $n+1$ down to n is equal to the long-run average rate at which it jumps up from n to $n+1$. Let π_n^{a} denote the fraction of arrivals finding n customers in the system, and let π_n^{d} be the fraction of departures leaving behind n customers in the system. The balance equation for the rates can be expressed as $\lambda\pi_n^{\mathrm{a}} = \lambda\pi_n^{\mathrm{d}}$, yielding $\pi_n^{\mathrm{a}} = \pi_n^{\mathrm{d}}$. By the PASTA property, the latter argument implies that the number of customers in steady state has the same distribution as the number of customers arriving during a sojourn time, so that $p_n = \pi_n^{\mathrm{d}}$. Now Proposition 1 immediately leads to (3).

We remark that, for $n = 0$, (2) yields the formula

$$p_0 = F_1(0) = \left(\sum_{n=0}^{\infty} \int_0^1 \frac{\mathrm{e}^{\lambda(1-u)}[-\lambda(1-u)]^n}{n!} \,\mathrm{d}G_n(u)\right)^{-1}, \tag{4}$$

whereas, by (3),

$$p_0 = \int_0^1 \mathrm{e}^{-\lambda x} \,\mathrm{d}(G * F_1)(x) + \mathrm{e}^{-\lambda}[1 - (G * F_1)(1)]. \tag{5}$$

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

Equation (5) has a probabilistic interpretation. Its right-hand side gives the probability that no arrival takes place during a steady-state sojourn time (split into the cases that this sojourn time is smaller than 1 or equal to 1), which is equal to π_0^d and thus to p_0. We do not know how to show the equivalence of (4) and (5) analytically.

3. Queueing model 2

We retain the notation of Section 2 for arrival epochs and service requirements. In queueing model 2 a customer whose waiting time in line would be greater than or equal to 1 is rejected. However, any customer whose waiting time would be smaller than 1 is fully served even though his/her sojourn time can be larger than 1. Accordingly, the associated workload process $\boldsymbol{V}^2 = \{V^2(t)\colon t \geq 0\}$ can be defined as follows:

$$V^2(t) = \begin{cases} \nu, & t = 0, \\ \max[V^2(T_{n-1}) - (t - T_{n-1}), 0], & T_{n-1} \leq t < T_n,\ n \geq 1, \\ V^2(T_n-) + S_n \mathbf{1}_{\{V^2(T_n-)<1\}}, & t = T_n,\ n \geq 1. \end{cases}$$

Thus, the system is blocked for new customers as long as its workload exceeds or is equal to 1. In the sequel we need a *modified version* $\tilde{\boldsymbol{V}}^2 = \{\tilde{V}^2(t)\colon t \geq 0\}$ of $\boldsymbol{V}^2$. The paths of $\tilde{\boldsymbol{V}}^2$ are obtained from those of $\boldsymbol{V}^2$ by deleting all idle periods and gluing together the busy periods.

We now consider $\tilde{\boldsymbol{V}}^2$ in the M/G/1 case in which the arrival times form a Poisson process of rate λ and the service requests have the common distribution function G. In this case $\tilde{\boldsymbol{V}}^2$ is a Markov process and, by level crossing theory [7], possesses a steady-state distribution $\tilde{F}_2$ with a density $\tilde{f}_2$, whose balance equation is given by

$$\begin{aligned} \tilde{f}_2(x) &= \lambda \int_0^{x\wedge 1} [1 - G(x-w)]\tilde{f}_2(w)\,\mathrm{d}w + \tilde{f}_2(0)[1 - G(x)] \\ &= \rho \int_0^{x \wedge 1} g_{\mathrm{e}}(x-w)\tilde{f}_2(w)\,\mathrm{d}w + b g_{\mathrm{e}}(x), \end{aligned} \tag{6}$$

where $g_{\mathrm{e}}(x) = \mu[1 - G(x)]$ is the equilibrium density associated with G, $\rho = \lambda/\mu$, and $b = \tilde{f}_2(0)/\mu$. Iterating (6) yields

$$\tilde{f}_2(x) = b \sum_{n=1}^{\infty} \rho^{n-1} g_{\mathrm{e}}^{*n}(x) = bQ(x), \qquad 0 \leq x \leq 1, \tag{7}$$

where g_{e}^{*n} is the n-fold convolution of g_{e} with itself and $Q(x) = \sum_{n=1}^{\infty} \rho^{n-1} g_{\mathrm{e}}^{*n}(x)$. To determine $\tilde{f}_2(x)$ for $x > 1$, we substitute (7) into (6) and get

$$\tilde{f}_2(x) = b g_{\mathrm{e}}(x) + b\rho \int_0^1 g_{\mathrm{e}}(x-w)Q(w)\,\mathrm{d}w, \qquad x > 1. \tag{8}$$

The constant b can be determined from the normalizing condition $\int_0^\infty \tilde{f}_2(w)\,\mathrm{d}w = 1$. We find that

$$b = \left[\int_0^1 Q(x)\,\mathrm{d}x + \int_1^\infty g_{\mathrm{e}}(x)\,\mathrm{d}x + \rho \int_1^\infty \int_0^1 g_{\mathrm{e}}(x-w)Q(w)\,\mathrm{d}w\,\mathrm{d}x\right]^{-1}. \tag{9}$$

The density $\tilde{f}_2$ is now given by (7)–(9).

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

The ordinary Laplace transform (LT) of $\tilde{f}_2$ can be expressed in terms of F_1 and p_0, which are given by (2) and (4). Let $\tilde{V}^1$ be the modified workload process of queueing model 1 in the M/G/1 case for which the idle periods are deleted and the busy periods are glued together. The steady-state distribution function $\tilde{F}_1(x)$ of $\tilde{V}^1$ is

$$\tilde{F}_1(x) = \frac{F_1(x) - p_0}{1 - p_0}. \tag{10}$$

Note that the busy cycles of both $\tilde{V}^1$ and $\tilde{V}^2$ are also their busy periods; assume that the initial cycles, $\tilde{B}_1$ and $\tilde{B}_2$, respectively, are started by a customer arriving at an empty system. A comparison of sample paths yields, for all $x \in [0, 1]$,

$$\int_0^{\tilde{B}_1} \mathbf{1}_{\{\tilde{V}^1(t) \le x\}}\,\mathrm{d}t = \int_0^{\tilde{B}_2} \mathbf{1}_{\{\tilde{V}^2(t) \le x\}}\,\mathrm{d}t. \tag{11}$$

Take expectations in (11), divide both sides by $\mathbb{E}\tilde{B}_1$, and take derivatives. For the steady-state densities $\tilde{f}_1$ and $\tilde{f}_2$ of $\tilde{V}^1$ and $\tilde{V}^2$, this yields the relation

$$\tilde{f}_2(x) = \frac{\mathbb{E}\tilde{B}_1}{\mathbb{E}\tilde{B}_2}\tilde{f}_1(x), \qquad 0 \le x \le 1.$$

On the other hand, by (10),

$$\tilde{f}_1(x) = \frac{1}{1 - p_0}\frac{\mathrm{d}}{\mathrm{d}x}F_1(x) = \frac{f_1(x)}{1 - p_0}, \qquad 0 \le x \le 1, \tag{12}$$

where f_1 denotes the density of F_1. Thus, the balance equation (6) can be written as

$$\tilde{f}_2(x) = \frac{\mathbb{E}\tilde{B}_1}{\mathbb{E}\tilde{B}_2}\biggl[\lambda \int_0^x [1 - G(x - w)]\tilde{f}_1(w)\,\mathrm{d}w + \tilde{f}_1(0)[1 - G(x)]\biggr]$$

for all $x \ge 0$; note that $\tilde{f}_1(x) = 0$ for $x > 1$. Taking LTs we obtain

$$\tilde{f}_2^*(\alpha) = \frac{\mathbb{E}\tilde{B}_1}{\mathbb{E}\tilde{B}_2}\biggl[\frac{\lambda}{\alpha}[1 - G^*(\alpha)]\tilde{f}_1^*(\alpha) + \frac{\tilde{f}_1(0)[1 - G^*(\alpha)]}{\alpha}\biggr].$$

Here and in the following f_i^* and $\tilde{f}_i^*$, $i = 1, 2$, denote the (ordinary) LTs of f_i and $\tilde{f}_i$. Since, by level crossing theory,

$$\mathbb{E}\tilde{B}_i = \frac{1}{\tilde{f}_i(0)}, \qquad i = 1, 2,$$

we arrive at

$$\tilde{f}_2^*(\alpha) = \tilde{f}_2(0)\biggl[\frac{1}{\tilde{f}_1(0)}\frac{\lambda}{\alpha}[1 - G^*(\alpha)]\tilde{f}_1^*(\alpha) + \frac{1 - G^*(\alpha)}{\alpha}\biggr], \tag{13}$$

where G^* is the LST of the distribution G. Now $\tilde{f}_1(0)$ is known from (12) at $x = 0$ and

$$\tilde{f}_1^*(\alpha) = \frac{1}{1 - p_0}\int_0^1 \mathrm{e}^{-\alpha x} f_1(x)\,\mathrm{d}x.$$

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

As $\tilde{f}_2^*(\alpha)$ is an LT, $\tilde{f}_2(0)$ can be obtained from the normalizing condition $\tilde{f}_2^*(0) = 1$. We obtain

$$\tilde{f}_2(0) = \frac{\mu \tilde{f}_1(0)}{\tilde{f}_1(0) + \lambda} = \frac{\mu f_1(0)}{(1 - p_0) f_1(0) + \lambda}.$$

Summarizing, we have proved the following result.

Theorem 1. *The steady-state density $\tilde{f}_2$ of $\tilde{\boldsymbol{V}}^2$ is given by $\tilde{f}_2(x) = bQ(x)$ for $x \in [0, 1]$ and*

$$\tilde{f}_2(x) = bg_{\mathrm{e}}(x) + b\rho \int_0^1 g_{\mathrm{e}}(x - w) Q(w) \,\mathrm{d}w \quad \text{if } x > 1,$$

where $Q(w)$ and b are defined in (7) and (9), respectively. An alternative formula is

$$\tilde{f}_2(x) = \frac{1}{(f_1(0) + \lambda)(1 - p_0)} \bigg[\lambda \int_0^x g_{\mathrm{e}}(x - w) f_1(w) \,\mathrm{d}w + f_1(0)[1 - G(x)] \bigg],$$

where $f_1 = F_1'$, the steady-state density of $\boldsymbol{V}^1$, is given by (2) and p_0 by (4) or (5). The LT $\tilde{f}_2^*$ of $\tilde{f}_2$ can be obtained in terms of the steady-state characteristics of $\boldsymbol{V}^1$ from (13).

Finally, let us return to the workload process $\boldsymbol{V}^2$.

Theorem 2. *The steady-state distribution function F_2 of $\boldsymbol{V}^2$ is given by*

$$F_2(x) = \pi_0 + (1 - \pi_0)\tilde{F}_2(x),$$

where $\pi_0 = \tilde{f}_2(0)/[\tilde{f}_2(0) + \lambda]$.

Proof. Clearly, π_0 is equal to the proportion of time queueing model 2 is idle. Since the expected values of the lengths of the idle and busy periods are $1/\lambda$ and $\mathbb{E}\tilde{B}_2 = 1/\tilde{f}_2(0)$, respectively, we have $\pi_0 = (1/\lambda)/[(1/\lambda) + (1/\tilde{f}_2(0))]$.

4. Duality analysis of queueing model 1 in the G/M/1 case

We now consider queueing model 1 with general interarrival distribution H and exponential service requests, i.e. $G(t) = 1 - \mathrm{e}^{-\lambda t}$. Clearly, the corresponding workload process $\boldsymbol{V}^1$ is regenerative. The duration of the first busy period is $B^1 = \inf\{t > 0 \colon V^1(t) = 0\}$ and that of the first busy cycle is $C^1 = \inf\{t > B \colon V^1(t) > 0\}$. The first idle period has length $I^1 = C^1 - B^1$.

Our aim is to derive the steady-state distributions of $\boldsymbol{V}^1$, I^1, B^1, and of the number of customers in the system via a duality with a certain Markovian system of the second type. This is carried out in two steps as follows.

1. We define a transformation that maps every path of $\boldsymbol{V}^1$ to a path of what will be seen as the *attained waiting time process* (AWT process) $\boldsymbol{A} = \{A(t) \colon t \geq 0\}$ of a certain queueing model of type 2. Define

$$A(t) = \begin{cases} 1 - V^1(t) & \text{if } V^1(t) > 0, \\ 1 + t - \inf\{s \in [0, t] \mid V^1(u) = 1 \text{ for all } u \in [s, t]\} & \text{if } V^1(t) = 0. \end{cases}$$

During the busy periods $\boldsymbol{A}$ is simply a reflection of $\boldsymbol{V}^1$ at the level $\frac{1}{2}$; during each idle period $\boldsymbol{A}$ increases linearly with slope 1 starting from level 1.

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

2. The *workload process* corresponding to the AWT process $\boldsymbol{A}$ is constructed as follows. Replace every negative jump of $\boldsymbol{A}$ by a linearly decreasing piece of trajectory with slope -1 on an interval whose length is equal to the negative jump size. Then replace the increasing pieces between the negative jumps of $\boldsymbol{A}$ by positive jumps whose sizes are equal to the corresponding linear increments. The process constructed in this way is called $\tilde{\boldsymbol{V}}^2$.

A little reflection shows there is an $\mathrm{M/G/1}$ queueing system of type 2 for which $A(t)$ can be interpreted as the time elapsed since the arrival of the customer who was last served before or is being served at time t, and $\tilde{\boldsymbol{V}}^2$ is the corresponding workload process. This underlying system has two types of customer, 'regular' and 'nonregular', and works as follows. Whenever the system becomes empty a new nonregular customer arrives instantaneously (so that there are no idle periods). The interarrival times between regular customers are exp(λ)-distributed, and all service times have distribution H. Customers who would have to wait in line for more than one time unit leave immediately without being served (so that the model is of type 2).

The negative jumps of $\boldsymbol{A}$ arrive according to a renewal process with interrenewal distribution H, and the jump sizes are exp(λ)-distributed and truncated at 0 when they would end below 0. In each period that $\boldsymbol{A}$ spends above 1 there is no jump, and this period is terminated by a downward jump such that the position after the jump can be written as $1 - \min[E_\lambda, 1]$, where E_λ is exponentially distributed and independent of the past evolution of $\boldsymbol{A}$.

A typical piece of a sample path of $\boldsymbol{A}$ is shown in Figure 1(b). It corresponds to that of $\boldsymbol{V}^{\mathbf{1}}$ in Figure 1(a), so that it has negative jumps at the times $A_1, A_2, \ldots,$ which are the departure times of the underlying (modified) queueing model 2. Applying the transformation described in step 2 above to the sample path of $\boldsymbol{A}$ in Figure 1(b) yields the workload sample path $\tilde{\boldsymbol{V}}^2$ depicted in Figure 1(c). The times $B_1, B_2, \ldots$ in Figure 1(b) and (c) are the arrival times of customers entering the system. Recall that in queueing model 2 customers do not enter if they have to wait in line more than 1 time unit. The times $V_1, V_2, \ldots$ are the arrival times of those customers who leave immediately without service due to this restriction. At time A_5 a busy period of queueing model 2 is terminated.

The importance of the AWT process is due the following theorem. It is the prototype of the kind of results that can be derived by the duality approach.

Theorem 3. *The steady-state laws of* $\tilde{\boldsymbol{V}}^2$ *and* $\boldsymbol{A}$ *are the same.*

Proof. The proof follows from the observation that, for $\tilde{\boldsymbol{V}}^2$ and $\boldsymbol{A}$, the hitting times of level 0 are the same and also the *peak* points (the local maximum points, which are the sojourn times) and the *trough* points (the local minimum points, which are the waiting times) between any consecutive hittings of 0 have the same values. Therefore, the rates of downcrossings of every level x coincide for $\tilde{\boldsymbol{V}}^2$ and $\boldsymbol{A}$, so that, by level crossing theory, their steady-state densities are equal (the steady-state distributions are absolutely continuous because the idle periods are deleted).

Theorem 3 enables us to compute the law of the $\mathrm{G/M/1}$ version of $\boldsymbol{V}^1$ with arrival rate λ and service rate μ from that of the $\mathrm{M/G/1}$ version of $\tilde{\boldsymbol{V}}^2$ with arrival rate μ and service rate λ. If we denote as before the corresponding steady-state densities by f_1 and $\tilde{f}_2$, respectively, we have

$$f_1(1-x) = \tilde{f}_2(x), \qquad x \in [0, 1);$$

$\tilde{f}_2$ was computed in Section 3.

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

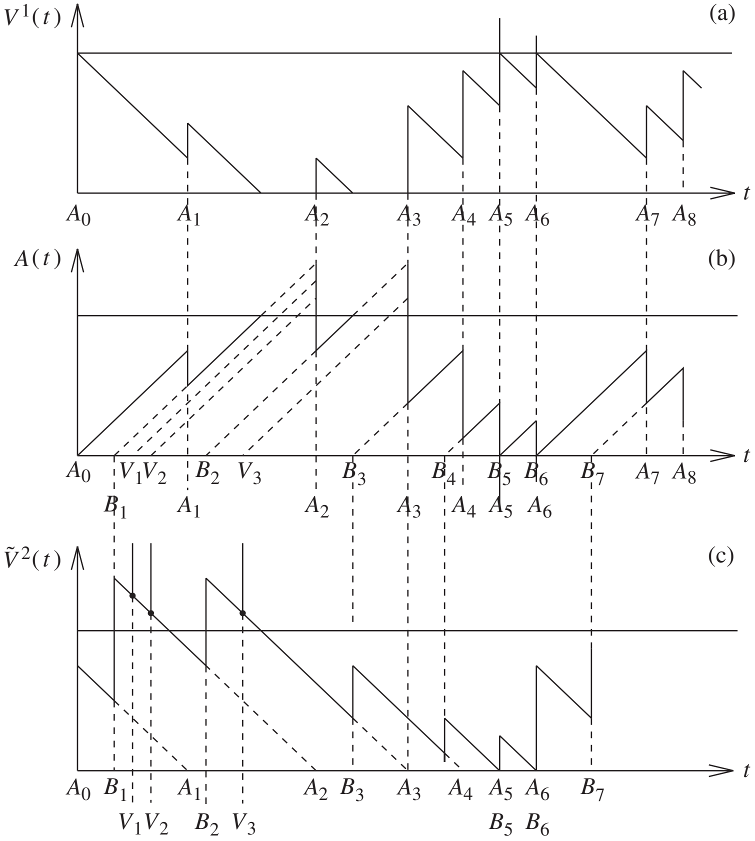

FIGURE 1: Typical sample paths of the processes $\boldsymbol{V}^1$, $\boldsymbol{A}$, and $\tilde{\boldsymbol{V}}^2$ connected by duality.

We now use the duality established above to derive several quantities of interest for the G/M/1 queueing model 1.

4.1. Idle and busy periods

Let $I^{\mathrm{G/M/1}}$ be the length of the idle period in steady state in queueing model 1. By duality, $I^{\mathrm{G/M/1}}$ can also be viewed as the overflow above level 1 in the modified version $\tilde{\boldsymbol{V}}^2$ (in which the idle periods are deleted) of queueing model 2 in the M/G/1 case. Obviously, overflows above 1 associated with $\tilde{\boldsymbol{V}}^2$ are i.i.d. random variables distributed as $I^{\mathrm{G/M/1}}$. The following theorem gives the LST of $I^{\mathrm{G/M/1}}$.

Theorem 4. *It holds that*

$$\mathbb{E}(\mathrm{e}^{-\alpha I^{\mathrm{G/M/1}}}) = 1 - \alpha \mathrm{e}^{\alpha} \frac{\int_1^\infty \mathrm{e}^{-\alpha x} \tilde{f}_2(x)\,\mathrm{d}x}{\tilde{f}_2(1)}.$$

Proof. The function defined by

$$\tilde{f}_2(x \mid x \geq 1) = \frac{\tilde{f}_2(x)}{\int_1^\infty \tilde{f}_2(w)\,\mathrm{d}w}, \qquad x \geq 1,$$

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

is the conditional equilibrium density of $\tilde{\boldsymbol{V}}^2$ given that it stays above level 1. Now we construct the process $\tilde{\boldsymbol{Y}} = \{\tilde{Y}(t)\colon t \geq 0\}$ from $\tilde{\boldsymbol{V}}^2$ by deleting the time periods in which $\tilde{\boldsymbol{V}}^2 \leq 1$ and gluing together the time periods in which $\tilde{\boldsymbol{V}}^2 > 1$. Clearly, $\tilde{f}_2(x \mid x \geq 1)$ is the steady-state density of $\tilde{\boldsymbol{Y}}$. Define $Y(t) = \tilde{Y}(t) - 1$. The process $\boldsymbol{Y} = \{Y(t)\colon t \geq 0\}$ can be interpreted as a forward recurrence time process with jump sizes distributed as $I^{\mathrm{G/M/1}}$; its equilibrium density is thus given by

$$f_{\mathrm{e}}(x) = \frac{\tilde{f}_2(x+1)}{\int_1^\infty \tilde{f}_2(w)\,\mathrm{d}w}, \qquad x \geq 0,$$

and f_{e} has the LT

$$f_{\mathrm{e}}^*(\alpha) = \frac{\int_0^\infty \mathrm{e}^{-\alpha x}\tilde{f}_2(x+1)\,\mathrm{d}x}{\int_1^\infty \tilde{f}_2(w)\,\mathrm{d}w} = \frac{\mathrm{e}^{\alpha}\int_1^\infty \mathrm{e}^{-\alpha x}\tilde{f}_2(x)\,\mathrm{d}x}{\int_1^\infty \tilde{f}_2(w)\,\mathrm{d}w}. \tag{14}$$

The process $\tilde{\boldsymbol{V}}^2$ is a regenerative process for which every downcrossing time of level x is a regeneration point. Choose $x = 1$. By interpreting the time $\tilde{\boldsymbol{V}}^2$ stays above level 1 as an *up* time and the time $\tilde{\boldsymbol{V}}^2$ stays below level 1 as a *down* time we obtain an alternating renewal process. Thus, the key renewal theorem yields

$$\int_1^\infty \tilde{f}_2(w)\,\mathrm{d}w = \frac{\mathbb{E}(I^{\mathrm{G/M/1}})}{\mathbb{E}(C)},$$

where C is a cycle length, defined as the time between two successive downcrossings of level 1. By level crossing theory,

$$\mathbb{E}(C) = \frac{1}{\tilde{f}_2(1)},$$

so that

$$\mathbb{E}(I^{\mathrm{G/M/1}}) = \frac{\int_1^\infty \tilde{f}_2(w)\,\mathrm{d}w}{\tilde{f}_2(1)}. \tag{15}$$

By standard renewal theory, the LST of the forward recurrence time is given by

$$f_{\mathrm{e}}^*(\alpha) = \frac{1 - \mathbb{E}(\mathrm{e}^{-\alpha I^{\mathrm{G/M/1}}})}{\alpha\mathbb{E}(I^{\mathrm{G/M/1}})}. \tag{16}$$

The theorem now follows from (14), (15), and (16).

The duality also allows us to deal with *busy periods* and *interoverflow times* of the G/M/1 models.

(i) By the construction above, we see that the length of a busy period in queueing model 1 of the G/M/1 type is stochastically equal to the time between two overflows in the modified queueing model 2 of the M/G/1 type in which the idle periods are deleted (see Figure 2). To determine this distribution, we need the solution of a two-sided first-exit problem for a standard M/G/1 workload process starting at level 1. When this process leaves (0, 1) for the first time by a jump above 1, this first exit time is equal to the first overflow time. When it exits at 0, we have to consider the next time instant it leaves (0, 1), and so on. We obtain a geometric series of i.i.d. exit times. An example of such a calculation is given in Section 6.

(ii) The time between two overflows in queueing model 1 of G/M/1 type is stochastically equal to the length of a busy period in queueing model 2 of M/G/1 type.

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

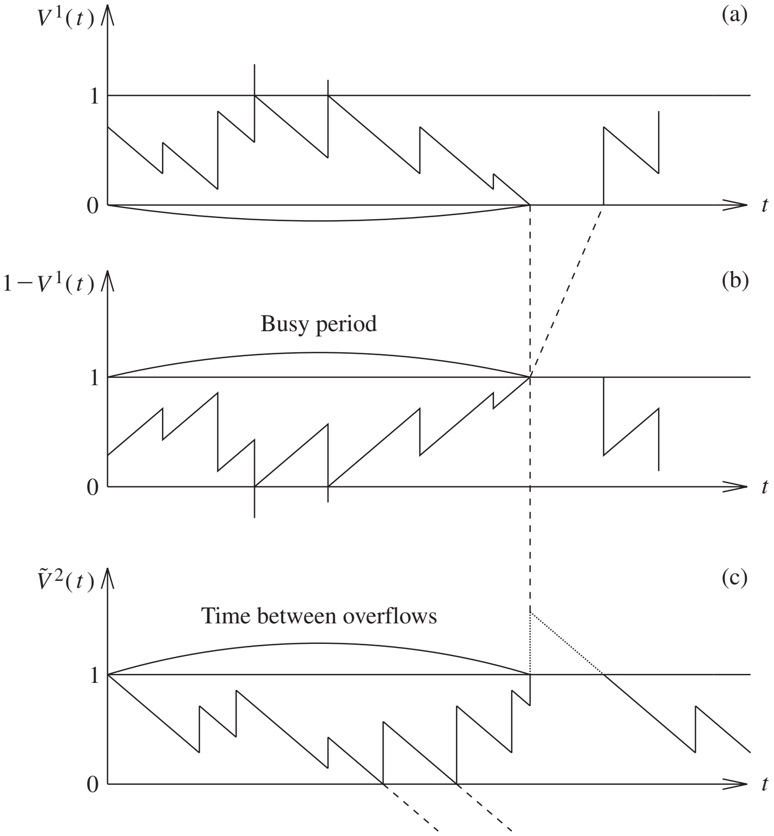

FIGURE 2: Duality between busy periods and interoverflow times. (a) A sample path of queueing model 1 (G/M/1 type). (b) A sample path generated by taking $1 - V^1(t)$. (c) A sample path generated by duality. The sample path in (c) describes queueing model 2 of M/G/1 type in which the idle periods are deleted.

4.2. The number of customers

The arrival times form a renewal process with interarrival distribution H. Thus, in steady state the probability that a departing customer leaves behind n customers is given by

$$p_n = \int_0^1 [H_n(x) - H_{n+1}(x)]\,\mathrm{d}F_{\mathrm{soj}}(x), \tag{17}$$

where H_n is the n-fold convolution of H with itself and $F_{\mathrm{soj}}(x)$ is the steady-state distribution function of the sojourn time.

The sojourn times are characterized by the heights of the *peak* points of the AWT process (the points just before negative jumps in Figure 1(b)). These peak points retain their values if the idle periods in the AWT process are deleted. Hence, F_{soj} is also the limiting distribution of the sojourn time in the modified process in which the idle periods are deleted and the busy periods are glued together. The fact that the arrival times of the negative jumps in the latter modified process form a Poisson process with rate μ enables us to apply the PASTA property to compute F_{soj} as follows. First we construct the process $\breve{\boldsymbol{A}} = \{\breve{A}(t) : t \ge 0\}$ by deleting the idle periods of $\boldsymbol{A}$ and gluing together the busy periods. We then define the process $\breve{\boldsymbol{V}} = \{\breve{V}(t) : t \ge 0\}$

by $\breve{V}(t) = 1 - \breve{A}(t)$. If $\breve{V}_{\mathrm{e}}$ is a random variable having the steady-state distribution of $\breve{\boldsymbol{V}}$, the limiting law of the sojourn time is the same as that of $1 - \breve{V}_{\mathrm{e}}$. The density $\breve{f}$ of $\breve{V}_{\mathrm{e}}$ satisfies the balance equation

$$\breve{f}(x) = \mu \int_0^x [1 - H(x - w)]\breve{f}(w)\,\mathrm{d}w + \breve{f}(0)[1 - H(x)], \qquad 0 \le x \le 1. \tag{18}$$

Solving for $\breve{f}(x)$ in (18) we obtain

$$\breve{f}(x) = ch_{\mathrm{e}} * \breve{K}(x),$$

where $h_{\mathrm{e}}(x) = (1 - H(x))/\int_0^\infty y\,\mathrm{d}H(y)$ is the equilibrium density corresponding to H,

$$\breve{K}(x) = \sum_{n=0}^{\infty} \rho^{-n} h_{\mathrm{e}}^{*n}(x),$$

and the constant c is computable from the normalizing condition $\int_0^1 \breve{f}(x)\,\mathrm{d}x = 1$, i.e.

$$c = \bigg[\int_0^1 \sum_{n=0}^{\infty} \rho^{-n} h_{\mathrm{e}}^{*n}(x)\,\mathrm{d}x\bigg]^{-1}.$$

It should be noted that the steady state density $\breve{f}$ is not just the derivative of the distribution F_{soj}, but F_{soj} can be expressed in terms of $\breve{f}$. To see this, recall first that the distribution F_{soj} has an atom at 1, since the sojourn time of some customers is truncated at level 1 if the waiting time plus their service requirement is greater than 1. This means that the negative jumps associated with $\breve{\boldsymbol{A}}$, or, alternatively, the positive jumps associated with $\breve{\boldsymbol{V}}$, are not Poisson jumps. In fact, the jumps of $\breve{\boldsymbol{V}}$ are generated by the composition of the Poisson arrivals (with rate μ), and the downcrossings of level 0 by $\breve{\boldsymbol{V}}$ (or level 1 by $\breve{\boldsymbol{A}}$). By the PASTA property, the Poisson jumps see the steady-state law of $\breve{\boldsymbol{V}}$, but the other jumps occur immediately after the sojourn times of the departing customers equal 1. By level crossing theory, the long-run average number of downcrossings of level 0 by $\breve{\boldsymbol{V}}$ is equal to $\breve{f}(0)$, so that the long-run average proportion of the Poisson jumps is $\mu/(\breve{f}(0) + \mu)$. From (17) and the discussion above we get the steady-state probability p_n that there are exactly n customers in the system:

$$p_n = \frac{\mu}{\mu + \breve{f}(0)} \int_0^1 [H_n(1 - x) - H_{n+1}(1 - x)]\breve{f}(x)\,\mathrm{d}x + \frac{\breve{f}(0)}{\mu + \breve{f}(0)}[H_n(1) - H_{n+1}(1)],$$

where $\breve{f}$ has been computed above.

5. M/G/1-type and G/M/1-type storage systems with a general outflow rule

We consider the content process $\boldsymbol{D}_{\mathrm{G/G/1}} = \{D_{\mathrm{G/G/1}}(t) \colon t \ge 0\}$ of a storage system with capacity 1, i.i.d. inputs whose arrival times form a renewal process, and a general outflow rule with rate function $r(\cdot)$. This storage system has been frequently studied in the case of Poisson arrivals; see [1], [8], [15], [16], [17], and [24]. The corresponding steady-state density $f_{\boldsymbol{D}_{\mathrm{M/G/1}}}$ of $\boldsymbol{D}_{\mathrm{M/G/1}}$ can be obtained from the Pollaczek–Khintchine balance equation

$$r(x) f_{\boldsymbol{D}_{\mathrm{M/G/1}}}(x) = \lambda \int_0^x [1 - G(x - w)] f_{\boldsymbol{D}_{\mathrm{M/G/1}}}(w)\,\mathrm{d}w + \lambda\pi(1 - G(x)), \qquad 0 < x \le 1,$$

where $\pi = \lim_{t\to\infty} \mathbb{P}(D_{\mathrm{M/G/1}}(t) = 0)$ is the atom of the steady-state distribution at 0 (see also [2, Section XIV.3]). Note that $\pi = 0$ if the release rate function is such that the storage system cannot be empty.

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

Our aim in this section is to derive the steady-state law of the content level for the analogous G/M/1-type storage system by generalizing the duality presented in Section 4. In the case of a general interarrival time distribution for the inputs there seems to be no direct way to compute this distribution.

We thus assume that the counting process $\boldsymbol{N} = \{N(t) \colon t \geq 0\}$ of the input arrivals is a renewal process whose interrenewal time distribution function G has finite mean $\mu^{-1} > 0$. The inputs (jump sizes) $Y_1, Y_2, \dots$ are independent and exponentially distributed with mean λ^{-1}. The storage system has finite capacity 1 so that the overflows above level 1 (which are also $\exp(\lambda)$-distributed) are lost. The output is governed by a general release rate function $r \colon [0, 1] \to [0, \infty)$, which is a positive, continuous function on $(0, 1]$ satisfying $r(0) = 0$. It follows from these conditions that $r(\cdot)$ is bounded away from 0 on compact intervals in $(0, 1]$; in particular,

$$0 < H(a, b) = \int_a^b \frac{\mathrm{d}u}{r(u)} < \infty \tag{19}$$

for $0 < a < b \leq 1$. The integral in (19) represents the time required to move from state b down to state a provided the input process is shut off (i.e. no jumps occur). We denote the content level process of this storage system by $\boldsymbol{D}_{\mathrm{G/M/1}} = \{D_{\mathrm{G/M/1}}(t) \colon t \geq 0\}$. We do not rule out the possibility that

$$H(0, x) = \int_0^x \frac{\mathrm{d}u}{r(u)} = \infty \tag{20}$$

for all $x > 0$. If (20) holds, state 0 is never reached from any initial state $x \in (0, 1]$. For example, the shot noise process has this property.

We now present a *sample path construction* of an M/G/1-type storage (content level) process which is 'dual' to the one above. Three stages are required.

(a) *The modified process* $\bar{\boldsymbol{D}}_{\mathrm{G/M/1}}$. The first auxiliary process (still of G/M/1 type) is $\bar{\boldsymbol{D}}_{\mathrm{G/M/1}} = \{\bar{D}_{\mathrm{G/M/1}}(t) \colon t \geq 0\}$, which differs from $\boldsymbol{D}_{\mathrm{G/M/1}}$ only during the dry periods. Let $T_0 = 0$, $T_1 = \inf\{t > 0 \colon D_{\mathrm{G/M/1}}(t) = 0\}$, and, for $n \geq 1$, recursively, $T_{n+1} = \inf\{t > T_n \colon D_{\mathrm{G/M/1}}(t) = 0 \text{ and } D_{\mathrm{G/M/1}}(s) > 0 \text{ for some } s \in (T_n, t)\}$. The last wet period completed before time t ends at $L(t) = T_{M(t)}$, where $M(t) = \max\{n \in \mathbb{Z}_+ \colon T_n < t\}$. The modified process $\tilde{\boldsymbol{D}}_{\mathrm{G/M/1}}$ is defined by

$$\bar{D}_{\mathrm{G/M/1}}(t) = \begin{cases} D_{\mathrm{G/M/1}}(t) & \text{if } 0 < D_{\mathrm{G/M/1}}(t) \leq 1, \\ -(t - L(t)) & \text{if } D_{\mathrm{G/M/1}}(t) = 0. \end{cases}$$

Thus, during the empty periods of the underlying storage system the process $\bar{\boldsymbol{D}}_{\mathrm{G/M/1}}$ is negative and decreases linearly with slope -1.

(b) *The risk process* $\boldsymbol{R}$. The risk process $\boldsymbol{R} = \{R(t) \colon t \geq 0\}$ is simply defined by $R(t) = 1 - \bar{D}_{\mathrm{G/M/1}}(t)$. Obviously, between its jumps $\boldsymbol{R}$ is governed by the input rate function

$$\bar{r}(u) = \begin{cases} r(1-u), & 0 < u < 1, \\ 1, & u \geq 1. \end{cases}$$

(c) *The Markovian storage process* $\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}$. Replace every negative jump of $\boldsymbol{R}$ by a linearly decreasing piece of trajectory with slope -1 on an interval whose length is equal to the negative jump size. Then, replace the increasing pieces between negative jumps of $\boldsymbol{R}$

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

by positive jumps whose sizes are equal to the differences between the corresponding *peak* points and *trough* points of $\boldsymbol{R}$. The resulting process is denoted by $\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}} = \{\tilde{D}_{\mathrm{M/G/1}}(t)\colon t \geq 0\}$. Note that, by this construction, corresponding peak points and trough points in $\boldsymbol{R}$ and $\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}$ have the same values.

The duality described here generalizes that introduced for $\boldsymbol{V}^1$ in Section 4. Note that while the jump sizes of $\tilde{\boldsymbol{V}}^2$ are i.i.d. random variables, those of the Markovian storage system $\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}$ are in general state dependent and are thus neither independent nor identically distributed. However, regardless of the laws of the jump sizes, it is clear that the long-run average rates of downcrossings of $\boldsymbol{R}$ and $\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}$ are equal.

The duality of $\boldsymbol{D}_{\mathrm{G/M/1}}$ and $\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}$ is expressed in the following relation of their steady-state laws. We have

$$r(1-x) f_{\boldsymbol{D}_{\mathrm{G/M/1}}}(1-x) = f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(x) \quad \text{for all } x \in (0,1)$$

and

$$\lim_{t\to\infty} \mathbb{P}(D_{\mathrm{G/M/1}}(t) = 0) = \int_1^\infty f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(x)\,\mathrm{d}x,$$

where $f_{\boldsymbol{D}_{\mathrm{G/M/1}}}$ and $f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}$ denote the steady-state densities of $\boldsymbol{D}_{\mathrm{G/M/1}}$ and $\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}$, respectively. Based on this relation we can now derive the steady-state law of $f_{\boldsymbol{D}_{\mathrm{G/M/1}}}$ and of $\boldsymbol{D}_{\mathrm{G/M/1}}$ using the Pollaczek–Khintchine-type equation for $\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}$ and the PASTA property.

Theorem 5. *It holds that*

$$f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(x) = \begin{cases} \dfrac{f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0)}{\lambda} K(x), & 0 < x \leq 1, \\ \dfrac{f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0)}{\lambda} \displaystyle\int_0^1 Q(x,w)K(w)\,\mathrm{d}w + \dfrac{f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0)}{\lambda} Q(x,0), & x > 1, \end{cases}$$

where

$$\begin{aligned} Q_1(x,w) &= Q(x,w) = 1 - G(H(w,x)), \\ Q_{n+1}(x,u) &= \int_u^x Q_n(x,w)Q(w,u)\,\mathrm{d}w, \qquad n \geq 1, \\ K(x) &= \sum_{n=1}^{\infty} Q_n(x,0), \end{aligned}$$

and

$$f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0) = \lambda \bigg(\int_0^1 K(x)\,\mathrm{d}x + \int_1^\infty \int_0^1 Q(x,w)K(w)\,\mathrm{d}w\,\mathrm{d}x + \int_1^\infty Q(x,0)\,\mathrm{d}x \bigg)^{-1}.$$

Proof. The balance equation of Pollaczek–Khintchine type for $f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}$ is given by

$$f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(x) = \begin{cases} \lambda \displaystyle\int_0^x [1 - G(H(w,x))] f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(w)\,\mathrm{d}w \\ \quad + f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0)[1 - G(H(0,x))], & 0 < x \leq 1, \\ \lambda \displaystyle\int_0^1 [1 - G(H(w,x))] f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(w)\,\mathrm{d}w \\ \quad + f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0)[1 - G(H(0,x))], & x > 1. \end{cases}$$

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

The density on the left-hand side is equal to the long-run average number of downcrossings of level x per unit time. Let us show that the right-hand side is equal to the corresponding long-run average number of upcrossings of x. Here λ is the arrival rate, and the integral is the probability that the state just before the jump is less than x and just after the jump is greater than x. Here we condition on $\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}$ being equal to w just before the jump; then we need the probability that a jump starting at w is of size greater than $x-w$. The time it takes for the risk process $\boldsymbol{R}$ to reach level x from level w is $H(w,x)$. For the jump to cross x, the interarrival time (with distribution function G) has to exceed $H(w,x)$ so that the probability in question is given by $1-G(H(w,x))$. The other term on the right-hand side covers the case that a jump from level 0 upcrosses x.

For $x \in (0,1]$, we have

$$\begin{aligned}
f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(x) &= \int_0^x Q(x,w)\biggl[\int_0^w Q(w,u) f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(u)\,\mathrm{d}u + \frac{f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0)}{\lambda} Q(w,0)\biggr]\mathrm{d}w \\
&\quad + \frac{f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0)}{\lambda} Q(x,0) \\
&= \int_0^x Q_2(x,u) f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(u)\,\mathrm{d}u + \frac{f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0)}{\lambda} Q_2(x,0) + \frac{f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0)}{\lambda} Q(x,0) \\
&= \cdots \\
&= \frac{f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0)}{\lambda} K(x).
\end{aligned}$$

It is not difficult to show that

$$0 \le Q_n(x,w) \le \frac{\lambda^n (x-w)^{n-1}}{(n-1)!},$$

so that the series for $K(x)$ is convergent and $\int_0^x Q_n(x,u) f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(u)\,\mathrm{d}u \to 0$ as $n \to \infty$. For $x > 1$, we obtain

$$f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(x) = \frac{f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0)}{\lambda} \int_0^1 Q(x,w)K(w)\,\mathrm{d}w + \frac{f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0)}{\lambda} Q(x,0).$$

We can now compute $f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(0)$ from the normalizing condition $\int_0^\infty f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(x)\,\mathrm{d}x = 1$. The proof is complete.

We conclude this section by stating an important relation for the peak points, i.e. local maxima, of the content process of the G/M/1-type storage system. Denote the successive heights of these peak points by $P_1, P_2, \dots$ and their steady-state density by $f_{\boldsymbol{PP}}(x)$.

Theorem 6. *The steady-state densities of* $P_1, P_2, \dots$ *and of* $\boldsymbol{D}_{\mathrm{G/M/1}}$ *are related as follows:*

$$f_{\boldsymbol{PP}}(x) = r(x) f_{\boldsymbol{D}_{\mathrm{G/M/1}}}(x), \qquad x \in (0,1). \tag{21}$$

Proof. The right-hand side of (21) is equal to the long-run average rate of downcrossings of level x by $\boldsymbol{D}_{\mathrm{G/M/1}}$. This rate is the same as the long-run average rate of upcrossings of level $1-x$ by the risk process $\boldsymbol{R}$ defined above, which in turn is equal to the long-run average rate of downcrossings of level $1-x$ by $\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}$. This latter downcrossing rate is given by $f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(1-x)$. By the PASTA property, $f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}$ is also the steady-state density of the sequence of trough points (local minima) of our special M/G/1-type storage system. But this is just the sequence $1-P_1, 1-P_2, \dots$. Therefore, $r(x) f_{\boldsymbol{D}_{\mathrm{G/M/1}}}(x) = f_{\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}}(1-x) = f_{\boldsymbol{PP}}(x)$.

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

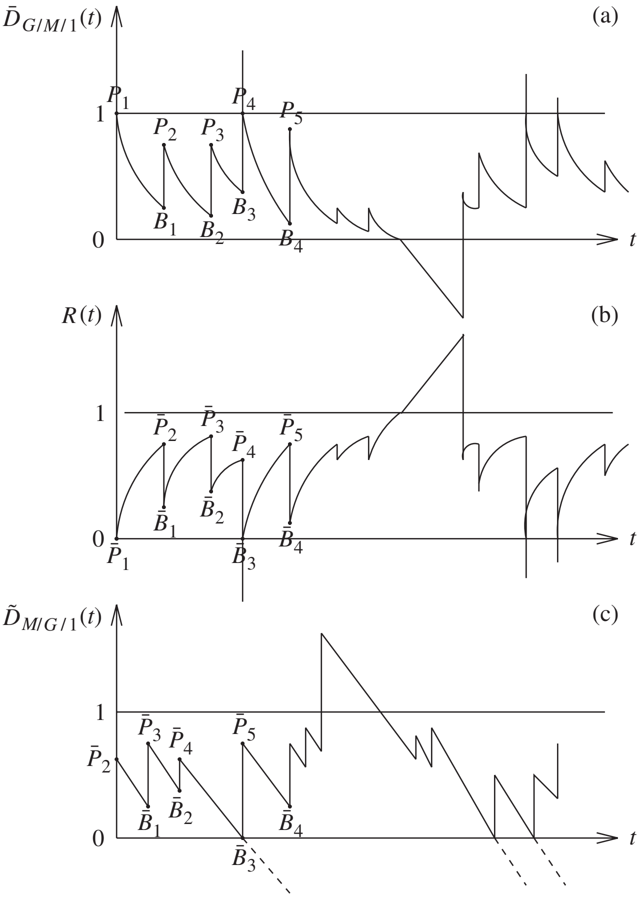

FIGURE 3: Typical sample paths of the processes $\bar{\boldsymbol{D}}_{\mathrm{G/M/1}}$, $\boldsymbol{R}$, and $\tilde{\boldsymbol{D}}_{\mathrm{M/G/1}}$ connected by the duality.

An illustration of the proof is given in Figure 3(a)–(c). The heights of the peak points $P_1, P_2, P_3, \dots$ and of the trough points $B_1, B_2, B_3, \dots$ in Figure 3(a) correspond to those in Figure 3(b) through the relations $\bar{P}_i = 1 - B_i$ and $\bar{B}_i = 1 - P_i$, and the same $\bar{P}_i$s and $\bar{B}_i$s appear in Figure 3(c).

Equation (21) relates the steady-state density of a continuous-time non-Markovian content process (of G/M/1 type) to that of a subsequence taken at discrete time instants (the peak points).

6. An example

The results for the G/M-type systems derived by the duality approach are explicit; however, the stationary densities are given in terms of infinite series of convolutions. In special cases more convenient formulae can be determined. We finally present an explicit computation: we determine the LST of the busy period of queueing model 1 in the $\mathrm{E}_2/\mathrm{M}/1$ case. The same derivation also works for $\mathrm{Erl}(k, \lambda)$, leading to k linear equations.

Consider queueing model 1 with $\mathrm{Erl}(2, 2\lambda)$ interarrival times (so that each exponential phase has rate 2λ and the arrival rate is λ) and $\exp(\mu)$-distributed service times. We know from Section 4 that the distribution of the busy period is the same as that of the time B between

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

two overflows in the corresponding $\tilde{\boldsymbol{V}}^2$ process of the modified queueing model 2 with Poisson arrivals of rate μ, service distribution function G having the density $g(x) = \mathrm{e}^{-2\lambda x}(2\lambda)^2 x$, and deleted idle periods.

We define the stopping time $\tau = \inf\{t > 0\colon \tilde{V}^2(t) \in \{0, 1\}\}$ and use the following notation:

$$\begin{aligned}
\Upsilon(\beta) &= \mathbb{E}\mathrm{e}^{-\beta B},\\
\psi_x(\beta) &= \mathbb{E}(\mathrm{e}^{-\beta B} \mid \tilde{V}^2(0) = x) = \mathbb{E}_x(\mathrm{e}^{-\beta B}),\\
\phi_*(x; \beta) &= \mathbb{E}_x(\mathrm{e}^{-\beta\tau}\mathbf{1}_{\{\tilde{V}^2(\tau)=0\}}),\\
\phi^*(x; \beta) &= \mathbb{E}_x(\mathrm{e}^{-\beta\tau}\mathbf{1}_{\{\tilde{V}^2(\tau)=1\}}),\\
\Psi(\beta) &= \int_0^1 \psi_x(\beta)\,\mathrm{d}G(x) = \int_0^1 \mathbb{E}_x(\mathrm{e}^{-\beta B})\mathrm{e}^{-\lambda x}\lambda^2 x\,\mathrm{d}x,\\
\Phi_*(\beta) &= \int_0^1 \mathbb{E}_x(\mathrm{e}^{-\beta\tau}\mathbf{1}_{\{\tilde{V}^2(\tau)=0\}})\mathrm{e}^{-\lambda x}\lambda^2 x\,\mathrm{d}x,\\
\Phi^*(\beta) &= \int_0^1 \mathbb{E}_x(\mathrm{e}^{-\beta\tau}\mathbf{1}_{\{\tilde{V}^2(\tau)>1\}})\mathrm{e}^{-\lambda x}\lambda^2 x\,\mathrm{d}x.
\end{aligned}$$

The above functions are connected by the renewal-type equation

$$\psi_x(\beta) = \phi^*(x; \beta) + \phi_*(x; \beta)[\mathrm{e}^{-2\lambda} + 2\lambda\mathrm{e}^{-2\lambda}] + \phi_*(x; \beta)\Psi(\beta), \qquad 0 < x \le 1. \tag{22}$$

The first term on the right-hand side of (22) is the improper LST of B restricted to the set $\{\tilde{V}^2(\tau) = 1\}$. The second term is the improper LST of B restricted to the set $\{\tilde{V}^2(\tau) = 0$, the jump size at time τ is at least $1\}$; note that $1 - G(1) = \mathrm{e}^{-2\lambda} + 2\lambda\mathrm{e}^{-2\lambda}$. The third term is the contribution of the case in which $\tilde{V}^2(\tau) = 0$ and the instantaneous jump at time τ goes to some $y < 1$.

Multiplying both sides of (22) by $\mathrm{e}^{-2\lambda x}(2\lambda)^2 x$ and integrating over $[0, 1)$, we obtain

$$\Psi(\beta) = \Phi^*(\beta) + [\mathrm{e}^{-2\lambda} + 2\lambda\mathrm{e}^{-2\lambda}]\Phi_*(\beta) + \Phi_*(\beta)\Psi(\beta). \tag{23}$$

Solving for $\Psi(\beta)$ in (23) yields

$$\Psi(\beta) = \frac{\Phi^*(\beta) + [\mathrm{e}^{-2\lambda} + 2\lambda\mathrm{e}^{-2\lambda}]\Phi_*(\beta)}{1 - \Phi_*(\beta)}. \tag{24}$$

By (24) and (22),

$$\begin{aligned}
\psi_x(\beta) &= \phi^*(x; \beta) + \phi_*(x; \beta)[\mathrm{e}^{-2\lambda} + 2\lambda\mathrm{e}^{-2\lambda}]\\
&\quad + \phi_*(x; \beta)\frac{\Phi^*(\beta) + [\mathrm{e}^{-2\lambda} + 2\lambda\mathrm{e}^{-2\lambda}]\Phi_*(\beta)}{1 - \Phi_*(\beta)}.
\end{aligned} \tag{25}$$

In particular, note that the LST of B is given by

$$\begin{aligned}
\Upsilon(\beta) &= \psi_1(\beta)\\
&= \phi^*(1; \beta) + \phi_*(1; \beta)[\mathrm{e}^{-2\lambda} + 2\lambda\mathrm{e}^{-2\lambda}]\\
&\quad + \phi_*(1; \beta)\frac{\Phi^*(\beta) + [\mathrm{e}^{-2\lambda} + 2\lambda\mathrm{e}^{-2\lambda}]\Phi_*(\beta)}{1 - \Phi_*(\beta)}.
\end{aligned} \tag{26}$$

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

The functionals $\Phi^*(\beta)$ and $\Phi_*(\beta)$ are simple integrals containing $\phi^*(x;\beta)$ and $\phi_*(x;\beta)$, so we only need to compute the latter functionals. To determine $\phi^*(x;\beta)$ and $\phi_*(x;\beta)$, we use a suitable Kella–Whitt martingale [18]. Take $Z(t)=\tilde{V}^2(t)+(\beta/\alpha)t$ for fixed $\beta>0$. Then

$$\tilde{M}(t)=(\tilde{\varphi}(\alpha)-\beta)\int_0^t \mathrm{e}^{-\alpha Z(u)}\,\mathrm{d}u+\mathrm{e}^{-\alpha Z(0)}-\mathrm{e}^{-\alpha Z(t)} \tag{27}$$

is a martingale for $t\le\tau$, where the exponent for the case of Erl$(2,2\lambda)$ jumps is given by

$$\tilde{\varphi}(\alpha)=\alpha-\mu\biggl[1-\biggl(\frac{2\lambda}{2\lambda+\alpha}\biggr)^2\biggr].$$

Applying the optional sampling theorem to the stopping time τ we obtain the equation

$$\biggl[\alpha-\mu\biggl(1-\biggl(\frac{2\lambda}{2\lambda+\alpha}\biggr)^2\biggr)-\beta\biggr]\mathbb{E}\biggl(\int_0^\tau \mathrm{e}^{-\alpha\tilde{V}^2(u)-\beta u}\,\mathrm{d}u\biggr)=-\mathrm{e}^{-\alpha a}+\mathbb{E}(\mathrm{e}^{-\alpha\tilde{V}^2(\tau)-\beta\tau}). \tag{28}$$

For fixed $\beta>0$, the function $\tilde{\varphi}(\alpha)-\beta$ has three real roots $\alpha_i(\beta)$, $i=1,2,3$, which satisfy

$$\alpha_1(\beta)>0>\alpha_2(\beta)>-2\lambda>\alpha_3(\beta). \tag{29}$$

The $\alpha_i(\beta)$ are the match points of the function $(2\lambda/(2\lambda+\alpha))^2$ and the linear function $1+\beta/\mu-\alpha/\mu$; it is easily seen that there are exactly three such points and that they satisfy (29). Next, from the memoryless property, either $\tilde{V}^2(\tau)-1\backsim$ Erl$(2,2\lambda)$ if level 1 is upcrossed by the first phase of the jump, or $\tilde{V}^2(\tau)-1\backsim\exp(\lambda)$ if level 1 is upcrossed by the second phase of the jump. Let F_1 (F_2) be the event that level 1 is upcrossed by the first (second) phase of the jump. Then, for $0<x\le1$,

$$\mathbb{E}_x(\mathrm{e}^{-\alpha V_1(\tau)-\beta\tau})=\theta_0(x;\beta)+\mathrm{e}^{-\alpha}\biggl(\frac{2\lambda}{2\lambda+\alpha}\biggr)^2\theta_1(x;\beta)+\mathrm{e}^{-\alpha}\frac{2\lambda}{2\lambda+\alpha}\theta_2(x;\beta),$$

where

$$\theta_1(x;\beta)=\mathbb{E}_x(\mathrm{e}^{-\beta\tau}\mathbf{1}_{F_1}),\qquad \theta_2(x;\beta)=\mathbb{E}_x(\mathrm{e}^{-\beta\tau}\mathbf{1}_{F_2}).$$

Thus, the left-hand side of (28) yields the three linear equations

$$\begin{aligned}\mathrm{e}^{-\alpha_i(\beta)x}&=\phi_*(x;\beta)+\mathrm{e}^{-\alpha}\biggl(\frac{2\lambda}{2\lambda+\alpha_i(\beta)}\biggr)^2\theta_1(x;\beta)\\&\qquad+\mathrm{e}^{-\alpha}\frac{2\lambda}{2\lambda+\alpha_i(\beta)}\theta_2(x;\beta),\qquad i=1,2,3,\end{aligned} \tag{30}$$

for the three unknowns $\theta_1(x;\beta)$, $\theta_2(x;\beta)$, and $\phi_*(x;\beta)$. The solution of this set of equations is available via computer algebra. However, the explicit formulae are lengthy and not very illuminating. Finally, by definition,

$$\phi^*(x;\beta)=\theta_1(x;\beta)+\theta_2(x;\beta).$$

Thus, in view of (25) and (26), we have all the components of $\Upsilon(\beta)$.

Remark. In the above computation we have inserted the roots $\alpha_i(\beta)$ in (28). However, (28) is derived from (27) and, thus, at first glance, only valid for $\operatorname{Re}\alpha>0$. This difficulty can, however, be overcome by analytic continuation. The function $\alpha\mapsto\tilde{\varphi}(\alpha)-\beta$ is analytic on $\mathbb{C}\setminus\{-2\lambda\}$. The function $\alpha\mapsto\mathbb{E}(\int_0^\tau \mathrm{e}^{-\alpha V_4(u)-\beta u}\,\mathrm{d}u)$ is easily seen to be analytic for all $\alpha\in\mathbb{C}$. Thus, after multiplication by $2\lambda+\alpha$, (30) becomes an identity between analytic functions which holds for $\operatorname{Re}\alpha>-2\lambda$ and, thus, by *analytic continuation*, for all $\alpha\in\mathbb{C}\setminus\{-2\lambda\}$. In particular, we may insert $\alpha=\alpha_i(\beta)$ into (28) so that (30) is valid.

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

7. Concluding remarks

In this study we have presented a duality that allows us to derive the characteristics of G/M/1-type models with restrictions on the accessibility from those of their 'dual' M/G/1-type counterparts with different restrictions, for which their Markovian structure can be exploited. We have considered the two most prominent such systems (truncated service time policy and bounded waiting time policy) and derived for the G/M/1 version of the first system the steady-state distributions of the workload, the number of customers present in the system, of busy and idle periods, and of the times between overflows via its duality with a restricted M/G/1 version of the second type of system. Other types of access restrictions require modifications in the duality constructions which may make the computations for the dual Markovian systems difficult (or even impossible). For example, in the case of quasirestricted accessibility the part of an incoming service requirement above the threshold is not totally refused but only a certain (random or deterministic) fraction is added to the workload. The G/M/1 version of this system has a dual modified M/G/1 system of our type 2 in which at the end of every busy period a new one is initiated immediately by a new customer whose service time distribution is that of the corresponding fraction of an exponential random variable. This dual system seems amenable to an analysis similar to the one given in this paper. Another (more challenging) problem is the generalization of the duality approach to random thresholds. As mentioned in the introduction, queueing model 2 can also be viewed as the customer impatience systems M/G/1 + D or G/M/1 + D, and it would be interesting to treat the case of random impatience via duality.

Queueing model 2 also appears in perishable inventory systems, where the process $\tilde{V}^2$ (in which the idle periods are deleted) can be interpreted as the so-called virtual outdating process of the stored perishable items (see the inventory model in [22]). In the latter case, if the value of an item is not constant over time but depends on its age via a certain function, the constant release rule of queueing model 2 has to be replaced by a general release rule [22]. The duality methods seem to be applicable to this model.

Acknowledgements

The authors thank Onno Boxma for his careful reading of the manuscript and his helpful comments and constructive suggestions. We are indebted to an anonymous referee whose thoughtful remarks led to a significantly improved presentation. Wolfgang Stadje was supported by the Deutsche Forschungsgemeinschaft.

References

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

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

[3] Bekker, R. and Zwart, B. (2005). On an equivalence between loss rates and cycle maxima in queues and dams. *Prob. Eng. Inf. Sci.* **19,** 241–255.

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

[5] Boxma, O., Perry, D., Stadje, W. and Zacks, S. (2009). The M/G/1 queue with quasi-restricted accessibility. *Stoch. Models* **25,** 151–196.

[6] Brandt, A. and Brandt, M. (2002). Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s + GI$ system. *Queueing Systems* **41,** 73–94.

[7] Brill, P. H. (2008). *Level Crossing Methods in Stochastic Models*. Springer, New York.

[8] Chen, H. and Yao, D. D. (1992). A fluid model for systems with random disruptions. *Operat. Res.* **S40,** S239–S247.

[9] Cohen, J. W. (1969). Single server queues with restricted accessibility. *J. Eng. Math.* **3,** 265–285.

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press

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

[11] Conolly, B. W., Parthasarathy, P. R. and Selvaraju, N. (2002). Double-ended queues with impatience. *Comput. Operat. Res.* **29,** 2053–2072.

[12] Daley, D. J. (1964). Single-server queueing systems with uniformly limited queueing times. *J. Austral. Math. Soc.* **4,** 489–505.

[13] De Kok, A. G. and Tijms, H. C. (1985). A queueing system with impatient customers. *J. Appl. Prob.* **22,** 688–696.

[14] Gavish, B. and Schweitzer, P. J. (1977). The Markovian queue with bounded waiting time. *Manag. Sci.* **23,** 1349–1357.

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

[16] Kaspi, H. and Perry, D. (1989). On a duality between a non-Markov storage/production process and a Markovian dam process with state dependent input and output. *J. Appl. Prob.* **26,** 835–844.

[17] Kaspi, H., Kella, O. and Perry, D. (1996). Dam processes with state dependent batch sizes and intermittent production processes with state dependent rates. *Queueing Systems* **24,** 37–57.

[18] Kella, O. and Whitt, W. (1992). Useful martingales for stochastic storage processes with Lévy input. *J. Appl. Prob.* **29,** 396–403.

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

[20] Liu, L. and Kulkarni, V. G. (2008). Busy period analysis for M/PH/1 queues with workload dependent balking. *Queueing Systems* **59**, 37–51.

[21] Löpker, A. and Perry, D. (2010). The idle period of the finite G/M/1 queue with an interpretation in risk theory. *Queueing Systems* **64,** 395–407.

[22] Nahmias, S., Perry, D. and Stadje, W. (2004). Actuarial valuation of perishable inventory systems. *Prob. Eng. Inf. Sci.* **18,** 219–232.

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

[24] Perry, D. and Stadje, W. (2003). Duality of dams via mountain processes. *Operat. Res. Lett.* **31,** 451–458.

[25] Perry, D., Stadje, W. and Zacks, S. (2000). Busy period analysis for M/G/1 and G/M/1 type queues with restricted accessibility. *Operat. Res. Lett.* **27,** 163–174.

[26] Stanford, R. E. (1990). On queues with impatience. *Adv. Appl. Prob.* **22,** 768–769.

[27] Zwart, A. P. (2006). A fluid queue with a finite buffer and subexponential input. *Adv. Appl. Prob.* **32,** 221–243.

https://doi.org/10.1239/jap/1378401226 Published online by Cambridge University Press