

SPOR-Report 2003-17

**On an equivalence between loss rates  
and cycle maxima in queues and dams**

R. Bekker  
A.P. Zwart

SPOR-Report  
Reports in Statistics, Probability and Operations Research

Eindhoven, June 2003  
The Netherlands

SPOR-Report  
Reports in Statistics, Probability and Operations Research

Eindhoven University of Technology  
Department of Mathematics and Computing Science  
Probability theory, Statistics and Operations research  
P.O. Box 513  
5600 MB Eindhoven - The Netherlands

Secretariat: Main Building 9.10  
Telephone: + 31 40 247 3130  
E-mail: [wscosor@win.tue.nl](mailto:wscosor@win.tue.nl)  
Internet: <http://www.win.tue.nl/math/bs/cosor.html>

ISSN 1567-5211

# On an equivalence between loss rates and cycle maxima in queues and dams

Rene Bekker and Bert Zwart

CWI

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

Eindhoven University of Technology

Department of Mathematics & Computer Science

P.O. Box 513, 5600 MB Eindhoven, The Netherlands

## Abstract

We consider the loss probability of a customer in a single-server queue with finite buffer and partial rejection, and show that it can be identified with the tail of the distribution of the cycle maximum of the associated infinite-buffer queue. This equivalence is shown to hold for the  $GI/G/1$  queue, and for dams with state-dependent release rates. To prove this equivalence, we use a version of Siegmund duality for stochastically monotone recursions, developed by Asmussen and Sigman (1996). As an application, we obtain several exact and asymptotic results for the loss probability and extend Takács' formula for the cycle maximum in the  $M/G/1$  queue to dams with variable release rate.

*2000 Mathematics Subject Classification:* 60K25 (primary), 68M20, 90B22 (secondary).

*Keywords & Phrases:* queues, storage processes, partial rejection, loss probability, cycle maximum, Siegmund duality, asymptotic expansions.

## 1 Introduction

Queueing models with finite buffers are useful to model systems where losses are of crucial importance, as in inventory theory and telecommunications. Unfortunately, finite-buffer queues are often more difficult to analyze than their infinite-buffer counterparts. An important exception is the  $GI/G/1$  queue where the total amount of work is upper bounded by  $K$  and customers are rejected under the partial-rejection discipline. This means that if a customer's sojourn time would exceed  $K$ , the customer only receives a fraction of its service requirement to make its sojourn time equal to  $K$ . This model is also known as the finite dam; see Section 2 for a precise description of the dynamics of this queue.

We consider the probability  $P_K$  that a customer gets partially rejected when entering the system in steady state. It is readily seen that

$$P_K = \mathbb{P}(W^K + S > K), \tag{1}$$

with  $W^K$  being the steady-state waiting time, and  $S$  a generic service time. Thus, information about  $P_K$  can be derived from the distribution of  $W^K$ .

Cohen [10], Chapter III.6, analyzed the distribution of  $W^K$  in the case that both the interarrival times and service times have a rational Laplace transform. For the  $M/G/1$  queue with  $\rho < 1$ , the distribution of  $W^K$  can be written in an elegant form, i.e., in terms of the steady-state waiting-time distribution of the  $M/G/1$  queue with infinite buffer size. This result is already known since Takács [21]. Using this result, Zwart [22] showed that  $P_K$  can be identified with Takács' expression [21] for the tail distribution of the cycle-maximum in the  $M/G/1$  queue, i.e., it is shown in [22] that

$$P_K = \mathbb{P}(C_{\max} > K). \quad (2)$$

For the  $GI/G/1$  queue with light-tailed service times, Van Ommeren and De Kok [18] derived exact asymptotics for  $P_K$  as  $K \rightarrow \infty$ . From their main result, it immediately follows that

$$P_K \sim \mathbb{P}(C_{\max} > K), \quad (3)$$

as  $K \rightarrow \infty$ . This naturally leads to the conjecture that (2) can be extended to the  $GI/G/1$  queue. Unfortunately, the proof in [22] can not be extended from Poisson to renewal arrivals, as it relies on exact computations for both  $P_K$  and the distribution of  $C_{\max}$ .

This brings us to the main goal of this paper: Our aim is to show that (a suitable modification of) (2) is valid for a large class of queueing models. In particular, we establish this equivalence without the need to compute both sides of (2) separately. Instead, the proof method in the present paper relates the distribution of  $W^K + S$  to a first-passage probability, which is in turn related to the distribution of  $C_{\max}$ . We will also give another proof based on a regenerative argument.

Both proof techniques strongly rely on a powerful duality theory for stochastic recursions, which has been developed by Asmussen & Sigman [5], and dates back to Lindley [16], Loynes [17], and Siegmund [20]. For a recent textbook treatment, see Asmussen [8]. This type of duality, also known as Siegmund duality, relates the stationary distribution of a given model to the first passage time of another model, called the dual model. Thus, Siegmund duality provides the right framework for proving (2). In its simplest form, Siegmund duality yields the well-known relationship between waiting-time probabilities for infinite-buffer queues and ruin probabilities.

This paper is organized as follows. We treat the  $GI/G/1$  queue in Section 2. Section 3 extends the results of Section 2 to state-dependent service rates. The final result for this class of models is somewhat more complicated than (2). In both sections, we give two proofs. These two proofs lead to different identities in Section 3. In Section 4 we show that (2) is not only useful to derive new results for the loss probability  $P_K$ , but also for the distribution of  $C_{\max}$ . Our results in this section include (i) a much shorter proof of the light-tailed asymptotics for  $P_K$  derived in [18], (ii) asymptotics of  $P_K$  for heavy-tailed service times, and (iii) an extension of Takács' formula for  $\mathbb{P}(C_{\max} > \cdot)$  to  $M/G/1$  queues with state dependent release rates. Concluding remarks can be found in Section 5.

## 2 The $GI/G/1$ queue

In this section we consider the  $GI/G/1$  queue with partial rejection, which is also known as the finite  $GI/G/1$  dam. Before we present our main result, we first introduce some notation and give a detailed model description.

Let  $T_1, T_2, \dots$  be the interarrival times of the customers and denote the arrival epoch of the  $n$ -th customer after time 0 by  $\bar{T}_n$ , i.e.,  $\bar{T}_n = \sum_{k=1}^n T_k$ . Assuming that a customer enter the system at time 0, the workload process  $\{D(t), t \in \mathbb{R}\}$  is then defined recursively by, cf. [10],

$$D(t) = \max(\min(D(\bar{T}_k^-) + S_k, K) - (t - \bar{T}_k), 0), \quad t \in [\bar{T}_k, \bar{T}_{k+1}). \quad (4)$$

Note that the workload process  $\{D(t), t \in \mathbb{R}\}$  is regenerative, with customer arrivals into an empty system being regeneration points. Let a regeneration cycle start at time 0, and define the first return time to 0 by

$$\tau_0 := \inf\{t > 0 : D(t) \leq 0\}. \quad (5)$$

Furthermore, let  $C_{\max}$  be the cycle maximum of a regeneration cycle, or, more formally,

$$C_{\max} := \sup\{D(t), 0 \leq t \leq \tau_0\}. \quad (6)$$

From the workload process in the finite  $GI/G/1$  dam we construct a dual risk process  $\{R(t), t \in \mathbb{R}\}$ , as in [19], by defining

$$R(t) := K - D(t). \quad (7)$$

The risk process is also regenerative and regeneration points in the risk process correspond to downward jump epochs from level  $K$ . Hence,  $\tau_0$  can be alternatively defined by  $\tau_0 := \inf\{t > 0 : R(t) \geq K\}$ .

Recall that  $P_K$  is the probability that an arriving customer is (partially) rejected. The main result in this section is the following theorem.

**Theorem 2.1.** *For the  $GI/G/1$  queue we have*

$$P_K = \mathbb{P}(C_{\max} \geq K). \quad (8)$$

In the remaining part of this section we present two proofs of Theorem 2.1. In the first proof, to be presented in Subsection 2.1, we take a direct approach, using the representation  $P_K = \mathbb{P}(W^K + S > K)$  and the above-mentioned definition of the cycle maximum. Equivalence is then shown using the machinery developed in [5].

The second proof, given in Subsection 2.2, establishes a link between the loss rate and the cycle maximum using an insightful regenerative argument. In particular, we use the fact that the number of losses in a cycle, given that at least one loss occurs, is geometrically distributed. The main step in this approach is the computation of the success parameter of that distribution. This is again established by results in [5].

## 2.1 Direct approach

Consider one regeneration cycle and let a customer enter the system at time 0. Since the workload process has peaks at time epochs just after an arrival instant, we may write

$$\begin{aligned}\mathbb{P}(C_{\max} \geq K) &= \mathbb{P}(\exists n \leq \tau_0 : W_n + S_n \geq K) \\ &= \mathbb{P}(\exists n \leq \tau_0 : R(\bar{T}_n^-) - S_n \leq 0).\end{aligned}\tag{9}$$

Observe that the right-hand side of (9) corresponds to a hitting probability; this probability may be interpreted as the probability that, starting in state  $K$ , (9) may be interpreted as the probability that state 0 is reached before the risk process hits state  $K$  again. Note that the risk process embedded at points  $\bar{T}_n$  is also recursively defined by the inter-arrival times and the service requirements. These two observations allow us to rewrite this embedded process as a monotone stochastic recursion with two absorbing states (0 and  $K$ ): We define  $R_0 = K$ ,  $R_{n+1} = g(R_n, U_n)$ , where  $U_n := (S_{n+1}, T_n)$ , and

$$g(x, s, t) = \begin{cases} 0, & \text{if } x = 0 \text{ or if } x \in (0, K] \text{ and } s \geq x, \\ x - (s - t), & \text{if } x \in (0, K] \text{ and } s < x, \\ \infty, & \text{if } x > K. \end{cases}\tag{10}$$

Thus, we start our recursion with initial reserve  $K$ , after which it evolves as an unrestricted random walk, until it leaves  $(0, K]$ . Moreover, it is always checked ahead whether a downward jump will not cause a negative workload, leading to absorbing state 0.

Now, Example 4 of Asmussen and Sigman [5] gives the corresponding dual stochastic recursion  $\{V_n\}$  which is defined as  $V_{n+1} = f(V_n, S_{n+1}, T_n)$ , with

$$f(y, s, t) = \min(((y - t)^+ + s), K).\tag{11}$$

This recursion corresponds to the workload right *after* a jump, or the sojourn time, of a finite  $GI/G/1$  dam. Under i.i.d. assumptions,  $V_n$  weakly converges to a random variable  $V$  as  $n \rightarrow \infty$ , see for example Chapter III.6 in Cohen [10]. Let

$$\gamma(x, K) := \min\{n \geq 1 : R_0 = x, R_n \notin (0, K]\},$$

denote the first exit time of  $(0, K]$ . Then, Corollary 3.1 of [5] yields the following fundamental result.

$$\mathbb{P}(V \geq x) = \mathbb{P}(R(x) \leq 0) = \mathbb{P}(R_{\gamma(x, K)}(x) \leq 0).\tag{12}$$

Thus, the distribution of  $V$  can be written as a first-passage probability. Using (9) and taking  $x = K$  in (12), we have

$$\begin{aligned}\mathbb{P}(C_{\max} \geq K) &= \mathbb{P}(R_{\gamma(K, K)}(K) \leq 0) \\ &= \mathbb{P}(V \geq K).\end{aligned}$$

Hence,

$$P_K = \mathbb{P}(W^K + S \geq K) \equiv \mathbb{P}(V \geq K) = \mathbb{P}(C_{\max} \geq K),\tag{13}$$

which completes the proof.

## 2.2 Regenerative approach

Let  $L_K$  be the number of not fully accepted customers, and  $N_K$  be the total number of customer arrivals during a regeneration cycle. A basic regenerative argument yields

$$P_K = \frac{\mathbb{E}L_K}{\mathbb{E}N_K}. \quad (14)$$

The denominator follows easily by

$$\mathbb{P}(W^K = 0) = \frac{1}{\mathbb{E}N_K} \mathbb{E} \left[ \sum_{i=1}^{N_K} I(W_i^K = 0) \right] = \frac{1}{\mathbb{E}N_K}, \quad (15)$$

where  $I(\cdot)$  is the indicator function.

The numerator may be rewritten as follows

$$\begin{aligned} \mathbb{E}L_K &= \mathbb{E}[L_K I(L_K \geq 1)] \\ &= \mathbb{E}[L_K \mid L_K \geq 1] \mathbb{P}(L_K \geq 1) \\ &= \mathbb{E}[L_K \mid L_K \geq 1] \mathbb{P}(C_{\max} \geq K). \end{aligned} \quad (16)$$

Moreover, observe that whenever the workload reaches level  $K$  and a customer is (partially) rejected, the process continues from level  $K$  starting with a new inter-arrival time, which clearly is independent of the past. Then, the probability of an additional customer loss in the regeneration cycle is equal to the probability that the workload process reaches level  $K$  again before the end of the busy cycle. Denoting  $\tau_K := \inf\{t > 0 : D(t) \geq K \mid D(0) = K\}$ , this leads to

$$\begin{aligned} \mathbb{P}(L_K \geq n+1 \mid L_K \geq n) &= \mathbb{P}(\tau_K < \tau_0 \mid D(0) = K) \\ &:= 1 - q_K. \end{aligned} \quad (17)$$

Iterating this argument, we conclude that  $L_K \mid L_K \geq 1$  is geometrically distributed with success parameter  $1 - q_K$ . Since the expectation of such a geometric distribution equals  $1/(1 - q_K)$ , we have to show that  $q_K = \mathbb{P}(W^K > 0)$  to complete the proof.

To do so, we use a similar construction of the risk-type process  $\{R(t), t \in \mathbb{R}\}$  as we did in the first proof. Note that (17) corresponds to the probability that from initial level 0, the risk process reaches level 0 again before it hits level  $K$ . Again, this can be transformed into a monotone stochastic recursion with two absorbing barriers, 0 and  $K$ : Define  $R_{n+1} = g(R_n, S_{n+1}, T_n)$ , with

$$g(x, s, t) = \begin{cases} 0, & \text{if } x = 0 \text{ or if } 0 < x < s - t, \\ x - (s - t), & \text{if } 0 < s - t \leq x \leq K - t, \\ \infty, & \text{if } x + t > K. \end{cases} \quad (18)$$

Thus, starting from level 0,  $R_n$  evolves as an unrestricted random walk until it leaves  $(0, K]$ . Note that it is indeed checked ahead whether the workload increases above level  $K$  before the next downward jump.

Now, another example of Asmussen and Sigman [5] gives the dual stochastic recursion  $\{V_n\}$ . In particular, Example 3 of [5] gives the dual function

$$f(y, s, t) = (\min(y + s, K) - t)^+,$$

defining the dual recursion  $V_{n+1} = f(V_n, S_{n+1}, T_n)$ . This recursion corresponds to the workload right *before* a jump, or the waiting time, in a finite  $GI/G/1$  dam. Use Corollary 3.1 of [5] and take  $x = \epsilon > 0$  in (12) to show that

$$\begin{aligned} q_K &= \lim_{\epsilon \downarrow 0} \mathbb{P}(R_{\gamma(\epsilon, K)}(\epsilon) \leq 0) \\ &= \lim_{\epsilon \downarrow 0} \mathbb{P}(V \geq \epsilon) = \mathbb{P}(V > 0). \end{aligned} \tag{19}$$

Recall that the  $V_n$  corresponds to the waiting time of the  $n$ -th customer, and  $V$  thus represent the *waiting* time in steady state. Combining (14)–(16), and (19) completes the proof.

**Remark 2.1.** *Both proofs relied on computing the dual of a recursion driven by a specific function  $f(x, z)$ , which is monotone in  $x$  for every  $z$ . In general, the driving function  $f$  and its dual  $g$  are related by*

$$\begin{aligned} g(x, z) &= \inf\{y : f(y, z) \geq x\}, \\ f(y, z) &= \inf\{x : g(x, z) \geq y\}. \end{aligned}$$

We refer to [5] (in particular Equation (2.4) of [5]) for details.

### 3 Dams with state-dependent release rates

In this section we consider the  $GI/G/1$  dam with general release rate. We start with introducing some definitions and a description of the driving sequence of the queueing process. Next, we state the main result and give two proofs analogous to the proofs in Section 2.

Let the release rate be  $r(x)$  when the workload equals  $x$ . We assume that  $r(0) = 0$  and that  $r(\cdot)$  is strictly positive, is left-continuous, and has a strictly positive right limit on  $(0, \infty)$ .

Also, define

$$\theta(x) := \int_0^x \frac{1}{r(y)} dy, \quad 0 < x < \infty, \tag{20}$$

representing the time required for a workload  $x$  to drain in the absence of arrivals. We assume that  $\theta(x) < \infty$ ,  $0 < x < \infty$ , implying that state zero can be reached in a finite amount of time. This ensures that  $C_{\max}$  is well-defined. Note that  $\theta(\cdot)$  is strictly increasing and we can thus unambiguously speak of  $\theta^{-1}(t)$ . Similar to [13] and [19], we define

$$q(u, t) := \theta^{-1}(\theta(u) - t). \tag{21}$$

Then  $q(u, t)$  is the workload level at time  $t$  if we start from level  $u$  at time 0 and no arrivals have taken place in between.

Denote the workload process of the  $GI/G/1$  queue with finite buffer  $K$  and general release rate function  $r(\cdot)$  by  $\{D^{r(\cdot)}(t), t \in \mathbb{R}\}$ . Let  $T_0 = 0$  and  $D^{r(\cdot)}(0) = x$ . Between jump epochs, the workload process is defined recursively by, cf. [19],

$$D^{r(\cdot)}(t) = q(D^{r(\cdot)}(\bar{T}_k^-), t), \quad \bar{T}_k < t < \bar{T}_{k+1}, \tag{22}$$

and at the  $(k + 1)$ -th jump epoch after time 0

$$D^{r(\cdot)}(\bar{T}_{k+1}) = \min \left( q(D^{r(\cdot)}(\bar{T}_k), T_{k+1}) + S_{k+1}, K \right). \quad (23)$$

Define  $\tilde{r}(x) := r(K - x)$ , for  $0 \leq x \leq K < \infty$ , and let all random variables  $X^{r(\cdot)}, X^{\tilde{r}(\cdot)}$  correspond to the model with release rate  $r(x), \tilde{r}(x)$ , respectively, if the process is at level  $x$ . Similar to Section 2, we construct a dual risk process  $\{R^{\tilde{r}(\cdot)}(t), t \in \mathbb{R}\}$ , by taking  $R^{\tilde{r}(\cdot)}(t) = K - D^{r(\cdot)}(t)$ . In between (the downward) jumps, the newly defined risk process is governed by the input rate function  $\tilde{r}(x) = r(K - x)$ , and the process satisfies

$$\frac{dR^{\tilde{r}(\cdot)}(t)}{dt} = \tilde{r}(R^{\tilde{r}(\cdot)}(t)).$$

Also, the risk process starts at  $R^{\tilde{r}(\cdot)}(0) := K - D^{r(\cdot)}(0)$ , and if the risk process starts at  $y$  and no jumps occur for  $t$  time units, its value increases, similar to the decrease in the workload process, to

$$\tilde{q}(y, t) := \tilde{\theta}^{-1}(\tilde{\theta}(y) + t). \quad (24)$$

Here,  $\tilde{\theta}(x) := \int_0^x (r(y))^{-1} dy$  represents the time required to move from 0 to  $x$  in the absence of negative jumps (claims), with inverse  $\tilde{\theta}^{-1}(t)$ . Note that, for finite  $K$ ,  $\int_0^x (\tilde{r}(y))^{-1} dy < \infty$ , meaning that state zero can be reached in a finite amount of time and the cycle maximum is also well-defined in this case.

**Theorem 3.1.** *For the GI/G/1 queue with general release rate we have*

$$P_K^{r(\cdot)} = \mathbb{P}(C_{\max}^{\tilde{r}(\cdot)} \geq K), \quad (25)$$

or alternatively,

$$P_K^{r(\cdot)} = \frac{\mathbb{P}(W^{K, r(\cdot)} = 0)}{\mathbb{P}(W^{K, \tilde{r}(\cdot)} = 0)} \mathbb{P}(C_{\max}^{\tilde{r}(\cdot)} \geq K). \quad (26)$$

We use a direct approach to show (25), thereby extending the proof in Section 2.1. To show (26), we follow the lines of Section 2.2, using an insightful regenerative argument and noting that the number of losses in a cycle, given that at least one loss occurs, has a geometric distribution. Let us start with (25).

*Proof of (25).* Since we assumed that  $\theta(x) < \infty$  for all finite  $x$ , the workload process  $\{D^{r(\cdot)}(t), t \in \mathbb{R}\}$  is still regenerative with customer arrivals into an empty system as regeneration points. The observation that the workload process has peaks at epochs right after an arrival instant, together with (9) and the construction of the risk process, leads to

$$\mathbb{P}(C_{\max}^{r(\cdot)} \geq K) = \mathbb{P}(\exists n \leq \tau_0 : R^{\tilde{r}(\cdot)}(T_n^-) - S_n \leq 0). \quad (27)$$

The probability in (27) can be interpreted as the hitting probability that state 0 is reached before the risk process hits state  $K$  again, starting from state  $K$ . Define  $R_0 = K$  and  $R_{n+1}^{\tilde{r}(\cdot)} = g(R_n^{\tilde{r}(\cdot)}, S_{n+1}, T_n)$ , with

$$g(x, s, t) = \begin{cases} 0, & \text{if } x = 0 \text{ or if } x \in (0, K] \text{ and } s \geq x, \\ \tilde{\theta}^{-1}(\tilde{\theta}(x - s) + t), & \text{if } x \in (0, K] \text{ and } s < x, \\ \infty, & \text{if } x > K. \end{cases} \quad (28)$$

Following [5], we construct the dual function corresponding to the described risk process

$$f(y, s, t) = \min\left(\tilde{\theta}^{-1}(\tilde{\theta}(y) - t) + s, K\right),$$

and define  $\{V_n\}$  recursively by  $V_{n+1}^{\tilde{r}(\cdot)} = f(V_n^{\tilde{r}(\cdot)}, S_{n+1}, T_n)$ . This process corresponds to a  $GI/G/1$  queue with release rate  $\tilde{r}(x) = r(K - x)$  if the workload equals  $x$ , embedded at epochs right *after* a jump. Combining the duality (12) between storage and risk processes with the expression (1) for  $P_K$ , now completes the proof of (25):

$$\begin{aligned} \mathbb{P}(C_{\max}^{r(\cdot)} \geq K) &= \mathbb{P}(R_{\gamma(K,K)}(K)^{\tilde{r}(\cdot)} \leq 0) \\ &= \mathbb{P}(V^{\tilde{r}(\cdot)} \geq K) \\ &= \mathbb{P}(W^{K, \tilde{r}(\cdot)} + S \geq K) \\ &= P_K^{\tilde{r}(\cdot)}. \end{aligned}$$

□

Next we turn to (26), which we show following the lines of Subsection 2.2.

*Proof of (26).* As  $\theta(x) < \infty$ , the workload process is still regenerative, and we consider the total number of (partially) rejected customers during a regeneration cycle. We apply the same regenerative argument as in Section 2.2 and note that customers are rejected if and only if the process reaches level  $K$  before the end of the cycle (which happens with probability  $\mathbb{P}(C_{\max}^{r(\cdot)} \geq K)$ ). Moreover, after a customer rejection, the process continues from level  $K$ , starting with a new interarrival time. This implies that the probability of an additional customer loss is independent of the past, or equivalently, that  $K$  is also a regeneration point. Therefore, we may conclude that, given that at least one loss occurs and the process starts from level  $K$ , the additional number of customer rejections is geometrically distributed with success parameter  $1 - q_K := \mathbb{P}(\tau_K < \tau_0 \mid D^{r(\cdot)}(0) = K)$ . Thus, we have to show that  $q_K = \mathbb{P}(W^{K, \tilde{r}(\cdot)} > 0)$  and combine (14)–(17) to complete the proof.

We start with the construction of the risk process  $\{R^{\tilde{r}(\cdot)}(t), t \in \mathbb{R}\}$  defined at the beginning of the section. We rewrite  $1 - q_K$  as the probability that, starting from level 0, the risk process hits level 0 again before it reaches level  $K$ . Interpreting our process as a monotone stochastic recursion with two absorbing barriers, we define  $R_{n+1}^{\tilde{r}(\cdot)} = g(R_n^{\tilde{r}(\cdot)}, S_{n+1}, T_n)$ , with

$$g(x, s, t) = \begin{cases} 0, & \text{if } x = 0 \text{ or if } \tilde{\theta}(x) < \tilde{\theta}(s) - t, \\ \tilde{\theta}^{-1}(\tilde{\theta}(x) + t) - s, & \text{if } \tilde{\theta}(s) - t < \tilde{\theta}(x) < \tilde{\theta}(K) - t, \\ \infty, & \text{if } \tilde{\theta}(x) + t > \tilde{\theta}(K). \end{cases} \quad (29)$$

Again, using [5] it can be seen that the dual recursion is defined as  $V_{n+1}^{\tilde{r}(\cdot)} = f(V_n^{\tilde{r}(\cdot)}, S_{n+1}, T_n)$ , with

$$f(y, s, t) = \tilde{\theta}^{-1}(\tilde{\theta}(\min(y + s, K)) - t).$$

The recursion corresponds to the workload at time epochs right *before* a jump. As the speed of the server is determined by the general release function, this generally does not equal the waiting time.

Finally, using Corollary 3.1 of [5] once more, we obtain

$$\begin{aligned} q_K &= \lim_{\epsilon \downarrow 0} \mathbb{P}(R_{\gamma(\epsilon, K)}^{\tilde{r}(\cdot)}(\epsilon) \leq 0) \\ &= \lim_{\epsilon \downarrow 0} \mathbb{P}(V^{\tilde{r}(\cdot)} \geq \epsilon) = \mathbb{P}(W_k^{\tilde{r}(\cdot)} > 0). \end{aligned} \quad (30)$$

Hence, by combining (14)–(17), and (30) we also have shown the second part of the result.  $\square$

**Remark 3.1.** *The constant  $\mathbb{P}(W^{K, r(\cdot)} = 0)/\mathbb{P}(W^{K, \tilde{r}(\cdot)} = 0)$  in (25) can easily be interpreted. As the interarrival times in both systems follow the same distribution, using (15), the constant equals the ratio of the respective cycle lengths.*

**Remark 3.2.** *A sample path argument can also provide some intuitive insight into the equivalence of (25) and (26). First, the risk process  $\{R^{\tilde{r}(\cdot)}(t) \mid t \geq 0\}$  can easily be interpreted as the available buffer capacity of a dam with release rate  $r(x)$  when the content equals  $x$ . Second, to convert the risk process into a queueing process again, we use a reversibility argument, as in [2, 4]. The sample path of this queueing process can essentially be obtained by time-reversing the sample path of the risk process, resulting in a queueing process with service speed  $\tilde{r}(x)$  when the workload equals  $x$ .*

## 4 Applications

In this section we state some exact and asymptotic results for  $P_K$ , by applying results for  $C_{\max}$  which are available in the literature. This leads to both new, and more transparent proofs of existing results.

### 4.1 Exact expressions for $P_K$

In the literature, there are several studies devoted to the distribution of  $C_{\max}$  for a variety of queueing models. We refer to Asmussen [7] for a survey of these results. The  $M/G/1$  case has already been treated in Zwart [22]. Here, we give an analogous result for the  $GI/M/1$  queue.

**Corollary 4.1.** *Consider the finite  $GI/M/1$  dam with  $\rho < 1$  and service rate  $\mu$ . Then*

$$P_K = \frac{1}{H(K)},$$

where  $H(x), x \geq 0$ , is a function with Laplace-Stieltjes transform (LST)

$$\frac{1}{s - \mu(1 - \alpha(s))},$$

with  $\alpha(s)$  the LST of the interarrival time distribution.

*Proof.* The result follows immediately from Theorem 2.1 and formula (7.76) on p. 627 of Cohen [10] stating that, for the  $GI/M/1$  queue,

$$\mathbb{P}(C_{\max} > K) = \frac{1}{H(K)}$$

with  $H(x), x \geq 0$ , defined as above.  $\square$

## 4.2 Asymptotics

Van Ommeren & De Kok [18] derive asymptotics for  $P_K$  in the  $GI/G/1$  queue under light-tailed assumptions. They conclude, after a lengthy argument, that (under their assumptions)  $P_K \sim \mathbb{P}(C_{\max} > K)$ , where  $f(x) \sim g(x)$  denotes  $f(x)/g(x) \rightarrow 1$  as  $x \rightarrow \infty$ . Asymptotics for the latter quantity are due to Iglehart [15]: Under certain regularity conditions (see [15]), it holds that

$$\mathbb{P}(C_{\max} > K) \sim De^{-\gamma K}, \quad (31)$$

for certain positive constants  $D$  and  $\gamma$ . Using Theorem 2.1, the proof of the main result of [18] is now trivial: Just combine Theorem 2.1 with (31) to (re-)obtain

$$P_K \sim De^{-\gamma K}. \quad (32)$$

For more details concerning specific assumptions and expressions for  $\gamma$  and  $C$  we refer to [15] and/or [18].

We proceed by giving results for the heavy-tailed case: Consider again the  $GI/G/1$  queue, but assume now that service times belong to the subclass  $\mathcal{S}^*$  of the class of subexponential distributions (see, e.g., Embrechts *et al.* [12] for a definition). This class contains all heavy-tailed distributions of interest, such as the Pareto, lognormal, and certain Weibull distributions.

Asymptotics for the cycle maximum can be found in [6]. If we combine these asymptotics with Theorem 2.1 we obtain (with  $N$  being the number of customers served in one busy cycle in the infinite buffer version of the  $GI/G/1$  queue)

**Corollary 4.2.** *If  $\rho < 1$  and the service time  $S \in \mathcal{S}^*$ , then*

$$P_K \sim \mathbb{E}N\mathbb{P}(S > K).$$

Also, in case of Poisson arrivals, this result can be extended to queues with general service speeds, see [6] for details. Note that (25) and (26), combined with Remark 3.1, indeed result in the same asymptotics.

## 4.3 Poisson arrivals and Takács' formula

The equivalence in Theorem 2.1 can also be used the other way around: Given information of  $P_K$ , we derive a new identity for the distribution of  $C_{\max}$  for queues with general release rate. For the special  $M/G/1$  case, the distribution of  $C_{\max}$  is known through Takács' formula. We combine the results of Section 3 with an identity for  $P_K$  which is valid under the additional assumption of Poisson arrivals. Under this assumption, the distribution of the amount of work in the system found by a customer  $W^{K,r(\cdot)}$  satisfies the following *proportionality* result:

$$\mathbb{P}(W^{K,r(\cdot)} \leq x) = \frac{\mathbb{P}(W^{r(\cdot)} \leq x)}{\mathbb{P}(W^{r(\cdot)} \leq K)} \quad (33)$$

Here,  $W^{r(\cdot)}$  is the steady-state amount of work in the system with  $K = \infty$  (assuming it exists). For the  $M/G/1$  queue, this result is well-known; see for example Takács [21], Cohen [10] and Hooghiemstra [14]. For a rigorous proof of (33) in case of a general release rate, we refer to Asmussen [1], Chapter XIII, Example 5.1.

Writing  $1 - P_K = \mathbb{P}(W^{K, r(\cdot)} + S \leq K)$ , conditioning on  $S$ , applying (33), and deconditioning on  $S$  then results in

$$\begin{aligned} P_x^{r(\cdot)} &= 1 - \mathbb{P}(W^{x, r(\cdot)} + S \leq x) \\ &= \frac{\mathbb{P}(W^{r(\cdot)} + S > x) - \mathbb{P}(W^{r(\cdot)} > x)}{\mathbb{P}(W^{r(\cdot)} \leq x)}. \end{aligned} \quad (34)$$

Combining this result with (26) then gives the following corollary:

**Corollary 4.3.** *For the  $M/G/1$  queue with infinite buffer size and general release rate we have*

$$\mathbb{P}(C_{\max}^{r(\cdot)} > x) = \frac{\mathbb{P}(W^{K, \tilde{r}(\cdot)} = 0)}{\mathbb{P}(W^{K, r(\cdot)} = 0)} \frac{\mathbb{P}(W^{r(\cdot)} + S > x) - \mathbb{P}(W^{r(\cdot)} > x)}{\mathbb{P}(W^{r(\cdot)} \leq x)}.$$

This is an extension of the classical formula for the distribution of  $C_{\max}$  in the  $M/G/1$  queue, which is due to Takács [21] (see also Cohen [9], and Asmussen & Perry [3] for alternative proofs). His result can be easily recovered from Corollary 4.3, since, for the  $M/G/1$  queue, we have  $r(x) \equiv \tilde{r}(x) \equiv 1$ . This yields the well-known formula

$$\mathbb{P}(C_{\max} \leq x) = \frac{\mathbb{P}(W + S \leq x)}{\mathbb{P}(W \leq x)}. \quad (35)$$

Related results for first-exit probabilities, as well as expressions for the distribution of  $W^{r(\cdot)}$  in terms of Volterra functions can be found in Harrison & Resnick [13]. Although Corollary 4.3 does not give a very explicit formula for the distribution of  $C_{\max}$  in general, we expect that this representation may be useful to obtain asymptotics and/or bounds. Asymptotic results in the light-tailed case are hardly known; see Asmussen [6, 7].

## 5 Conclusion

We have considered several queueing models which operate under the partial-rejection mechanism. For these models, we have shown that the loss probability of a customer can be identified with the tail probability of the cycle maximum.

The present work raises several questions that could be interesting for further research. First of all, we believe that a suitable modification of Theorem 2.1 still holds for other queueing models, such as queueing models with Markov-modulated input. This is potentially useful, since the distribution of the cycle maximum is known for a large class of such models; see Asmussen & Perry [3].

Furthermore, we expect that Siegmund duality and related results can also be fruitful in other queueing problems. In the context of the present paper, we believe that an analogue of (2) can be shown for queues which can be modeled as birth-death processes: Siegmund-type duality results for birth-death processes have been derived by Dette *et al.* [11].

## References

- [1] Asmussen, S. (1987). *Applied Probability and Queues*. Wiley, New York.
- [2] Asmussen, S., Schok Petersen, S. (1988). Ruin probabilities expressed in terms of storage processes. *Advances in Applied Probability* **20**, 913–916.
- [3] Asmussen, S., Perry, D. (1992). On cycle maxima, first passage problems and extreme value theory for queues. *Stochastic Models* **8**, 421–458.
- [4] Asmussen, S., Kella, O. (1996). Rate modulation in dams and ruin problems. *Journal of Applied Probability* **33**, 523–535.
- [5] Asmussen, S., Sigman, K. (1996). Monotone stochastic recursions and their duals. *Probability in the Engineering and Informational Sciences* **10**, 1–20.
- [6] Asmussen, S. (1998). Subexponential asymptotics for stochastic processes: extremal behaviour, stationary distributions and first passage times. *Annals of Applied Probability* **8**, 354–374.
- [7] Asmussen, S. (1998). Extreme value theory for queues via cycle maxima. *Extremes* **2**, 137–168.
- [8] Asmussen, S. (2003). *Applied Probability and Queues*. Second edition. Springer, New York.
- [9] Cohen, J.W. (1976). *Regenerative Processes in Queueing Theory*. Springer, Berlin.
- [10] Cohen, J.W. (1982). *The Single Server Queue*. North Holland, Amsterdam.
- [11] Dette, H., Fill, J.A., Pitman, J., Studden, W.J. (1997). Wall and Siegmund duality relations for birth and death chains with reflecting barrier. *Journal of Theoretical Probability* **10**, 349–374.
- [12] Embrechts, P., Klüppelberg, C., Mikosch, T. (1997). *Modelling Extremal Events*. Springer, Berlin.
- [13] Harrison, J.M., Resnick, S.I. (1976). The stationary distribution and first exit probabilities of a storage process with general release rule. *Mathematics of Operations Research* **1**, 347–358.
- [14] Hooghiemstra, G. (1987). A path construction for the virtual waiting time of an  $M/G/1$  queue. *Statistica Neerlandica* **41**, 175–181.
- [15] Iglehart, D.G. (1972). Extreme values in the  $GI/G/1$  queue. *Annals of Mathematical Statistics* **43**, 627–635.
- [16] D.V. Lindley (1959). Discussion of a paper by C.B. Winsten. *Proceedings of the Cambridge Philosophical Society* **48**, 277–289.
- [17] Loynes, R.M. (1965). On a property of the random walks describing simple queues and dams. *Journal of the Royal Statistical Society Series B* **27**, 125–129.
- [18] Van Ommeren, J.C.W., De Kok, A.G. (1987). Asymptotic results for buffer systems under heavy load. *Probability in the Engineering and Informational Sciences* **1**, 327–348.
- [19] Perry, D., Stadje, W. (2003). Duality of dams via mountain processes. Technical report University of Haifa, Israel. Available at <http://www.mathematik.uni-osnabrueck.de/staff/phpages/stadjew/publications.html>
- [20] Siegmund, D. (1976). The equivalence of absorbing and reflecting barrier problems for stochastically monotone Markov processes. *Annals of Probability* **4**, 914–924.
- [21] Takács, L. (1967). *Combinatorial Methods in the Theory of Stochastic Processes*. Wiley, New York.
- [22] Zwart, A.P. (2000). A fluid queue with a finite buffer and subexponential input. *Advances in Applied Probability* **32**, 221–243.