Probability in the Engineering and Informational Sciences, **19**, 2005, 241–255. Printed in the U.S.A.

ON AN EQUIVALENCE BETWEEN LOSS RATES AND CYCLE MAXIMA IN QUEUES AND DAMS

RENÉ BEKKER AND BERT ZWART
CWI
1090 GB Amsterdam, The Netherlands
and
Department of Mathematics & Computer Science
Eindhoven University of Technology
5600 MB Eindhoven, The Netherlands
E-mail: rbekker@win.tue.nl; zwart@win.tue.nl

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 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 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.

1. INTRODUCTION

Queuing 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 rejection discipline operates such 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.

 0269-9648/05 $16.00

https://doi.org/10.1017/S0269964805050138 Published online by Cambridge University Press

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

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

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

Cohen [11, Chap. 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 traffic intensity $\rho < 1$, 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 [23] showed that P_K can be identified with Takács' formula [21] for the tail distribution of the cycle maximum in the $M/G/1$ queue; that is, it is shown in [23] that

$$P_K = \mathbb{P}(C_{\max} \geq K). \tag{2}$$

For the $GI/G/1$ queue with light-tailed service times, Van Ommeren and De Kok [22] derived exact asymptotics for P_K as $K \to \infty$. From their main result, it immediately follows that

$$P_K \sim \mathbb{P}(C_{\max} \geq K),$$

as $K \to \infty$. This naturally leads to the conjecture that (2) can be extended to the $GI/G/1$ queue. Unfortunately, the proof in [23] cannot 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 the article: Our aim is to show that (an appropriate modification of) (2) is valid for a large class of queuing models. In particular, we establish this equivalence for any positive ρ without the need to compute both sides of (2) separately. Instead, the proof method in the present article 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 and Sigman [7], and dates back to Lindley [17], Loynes [18], and Siegmund [20]. For a recent textbook treatment, see Asmussen [1]. 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 article is organized as follows. We treat the $GI/G/1$ queue in Section 2. Section 3 extends the results of Section 2 to queues with state-dependent service

https://doi.org/10.1017/S0269964805050138 Published online by Cambridge University Press

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 main results in this section are (1) a much shorter proof of the light-tailed asymptotics for P_K derived in [22], (2) asymptotics of P_K for heavy-tailed service times, and (3) 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 $T_1, T_2, \ldots$ be the interarrival times of customers and denote the nth arrival epoch after time 0 by $\overline{T}_n$ (i.e., $\overline{T}_n = \sum_{k=1}^{n} T_k$). We assume that the interarrival times form an independent and identically distributed (i.i.d.) sequence and that $\mathbb{E}[T_1] < \infty$. The service requirement of the nth customer is denoted by S_n, $n = 1, 2, \ldots$, whereas $S_1, S_2, \ldots$ are also assumed to be i.i.d. Define $\rho := \mathbb{E}[S_1]/\mathbb{E}[T_1]$ as the load of the system. We like to emphasize that ρ may take any (positive) value. To obtain a nontrivial model though, we additionally assume that $\mathbb{P}(T_1 > S_1) > 0$ (otherwise $P_K \equiv 1$).

The workload process $\{D(t), t \in \mathbb{R}\}$ is now defined recursively by (cf. [11])

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

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

$$\tau_0 := \inf\{t > 0 : D(t) \le 0\}.$$

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

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

Observe that, for $x \le K$, $\mathbb{P}(C_{\max} \ge x)$ is the same for the finite dam and its infinite-buffer counterpart. So, without affecting the results, we will henceforth adopt the above definition of $C_{\max}$ when we consider the cycle maximum in the $GI/G/1$ queue with infinite-buffer capacity. Note that $\mathbb{P}(C_{\max} = \infty) > 0$ if $K = \infty$ and $\rho > 1$.

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

$$R(t) := K - D(t).$$

https://doi.org/10.1017/S0269964805050138 Published online by Cambridge University Press

Away from the boundaries, this process increases linearly at rate 1 and negative jumps occur at times $\overline{T}_n$ of size S_n, $n = 1,2,\ldots$. By definition, jumps below zero are truncated and if the process hits state K, it remains in K until the next (downward) jump (see Fig. 1 for an illustration). In fact, we are only interested in the behavior of $R(t)$ until it hits one of the boundaries, and in this region, the process $\{R(t), t \in \mathbb{R}\}$ shows a strong resemblance to a risk process (where 0 is supposed to be an absorbing state).

Due to the finite capacity, the process $\{R(t), t \in \mathbb{R}\}$ is also regenerative, and regeneration points in the process correspond to downward jump epochs from level K. Hence, τ_0 can be alternatively defined by $\tau_0 := \inf\{t > 0 : R(t) \geq K\}$.

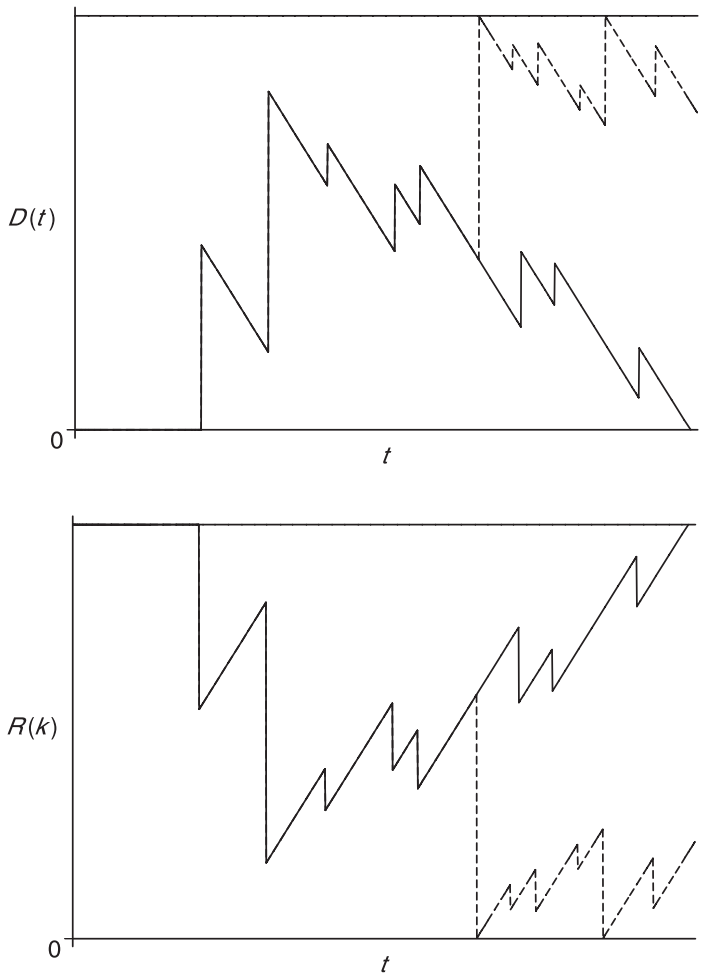

FIGURE 1. Two sample paths of $D(t)$ until it hits one of the boundaries, with corresponding $R(t)$.

https://doi.org/10.1017/S0269964805050138 Published online by Cambridge University Press

Recall that P_K is the steady-state 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*

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

In the remaining part of this section, we present two proofs of Theorem 2.1. In the first proof, to be presented in Section 2.1, we take a direct approach, using the representation $P_K = \mathbb{P}(W^K + S \geq K)$ and the above-mentioned definition of the cycle maximum. Equivalence is then shown using the machinery developed in [7].

The second proof, given in Section 2.2, establishes a link between the loss rate and the cycle maximum using an insightful regenerative argument. In particular, we utilize 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 [7].

2.1. Direct Approach

As mentioned, to determine the tail distribution of the cycle maximum in an infinite-capacity model, we may also assume that the workload is uniformly bounded as described earlier. So, consider one regeneration cycle of the process $\{D(t), t \in \mathbb{R}\}$ (or equivalently $\{R(t), t \in \mathbb{R}\}$) 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

$$\begin{aligned} \mathbb{P}(C_{\max} \geq K) &= \mathbb{P}(\exists n \leq \tau_0 : W_n + S_n \geq K) \\ &= \mathbb{P}(\exists n \leq \tau_0 : R(\overline{T}_n^-) - S_n \leq 0). \end{aligned} \tag{3}$$

Observe that the right-hand side of (3) corresponds to a hitting probability; starting in state K, (3) may be interpreted as the probability that state 0 is reached before $R(t)$ hits state K again. Note that the process $\{R(t), t \in \mathbb{R}\}$ embedded at points $\overline{T}_n$ is also recursively defined by the interarrival 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

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

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 absorption in state 0.

https://doi.org/10.1017/S0269964805050138 Published online by Cambridge University Press

Now, Example 4 of Asmussen and Sigman [7] gives the corresponding dual stochastic recursion $\{V_n\}$, which is defined as $V_{n+1} = f(V_n, S_{n+1}, T_n)$, where

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

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 $n \to \infty$; see, for example, Cohen [11, Chap. III.6]. 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 [7] yields the following fundamental result:

$$\mathbb{P}(V \geq x) = \lim_{n\to\infty} \mathbb{P}(R_n \leq 0 | R_0 = x) = \mathbb{P}(R_{\gamma(x,K)} \leq 0). \tag{4}$$

Thus, the distribution of V can be written as a first-passage probability. Using (3) and taking $x = K$ in (4), we have

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

Hence,

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

which completes the proof.

2.2. Regenerative Approach

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

$$P_K = \frac{\mathbb{E}L_K}{\mathbb{E}N_K}. \tag{5}$$

The denominator follows easily by

$$\mathbb{P}(W^K = 0) = \frac{1}{\mathbb{E}N_K}\,\mathbb{E}\left[\sum_{i=1}^{N_K} I(W_i^K = 0)\right] = \frac{1}{\mathbb{E}N_K}, \tag{6}$$

where $I(\cdot)$ is the indicator function. The numerator may be rewritten as

$$\begin{aligned}\mathbb{E}L_K &= \mathbb{E}[L_K I(L_K \geq 1)]\\ &= \mathbb{E}[L_K | L_K \geq 1]\mathbb{P}(L_K \geq 1)\\ &= \mathbb{E}[L_K | L_K \geq 1]\mathbb{P}(C_{\max} \geq K).\end{aligned} \tag{7}$$

https://doi.org/10.1017/S0269964805050138 Published online by Cambridge University Press

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 interarrival 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 $\tau_K := \inf\{t > 0 : D(t) \geq K | D(0) = K\}$, this leads to

$$\begin{aligned} \mathbb{P}(L_K \geq n+1 | L_K \geq n) &= \mathbb{P}(\tau_K < \tau_0 | D(0) = K) \\ &:= 1 - q_K. \end{aligned} \tag{8}$$

Iterating this argument, we conclude that $L_K | L_K \geq 1$ 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 (8) corresponds to the probability that from initial level 0, $R(t)$ 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

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

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 [7] provides the dual stochastic recursion $\{V_n\}$. In particular, Example 3 of [7] 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 [7] and take $x = \epsilon > 0$ in (4) to show that

$$\begin{aligned} q_K &= \lim_{\epsilon \downarrow 0} \mathbb{P}(R_{\gamma(\epsilon,K)} \leq 0) \\ &= \lim_{\epsilon \downarrow 0} \mathbb{P}(V \geq \epsilon) = \mathbb{P}(V > 0). \end{aligned} \tag{9}$$

Recall that the V_n corresponds to the waiting time of the nth customer, and V thus represents the *waiting* time in steady state. Combining (5)–(9) completes the proof.

Remark 2.1: Both proofs rely 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

https://doi.org/10.1017/S0269964805050138 Published online by Cambridge University Press

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

We refer to [7] (in particular, Eq. (2.4)) 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 queuing process. Next, we state the main result and give two proofs, analogous to the proofs in Section 2.

Consider the model of Section 2, but 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, left-continuous, and has a strictly positive right limit on $(0,\infty)$. Also, define

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

representing the time required for a workload x to drain in the absence of arrivals. We assume that $\theta(x) < \infty$, $0 < x < \infty$, indicating that state 0 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, unambigously speak of $\theta^{-1}(t)$. Similar to [14,19], we define

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

Then $q(u,t)$ represents the workload level at time t if we start off 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])

$$D^{r(\cdot)}(t) = q(D^{r(\cdot)}(\overline{T}_k^-), t), \qquad \overline{T}_k < t < \overline{T}_{k+1},$$

and at the $(k+1)$st jump epoch after time 0,

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

To exclude trivial cases where the workload is bounded from below, we assume that $\mathbb{P}(q(x + S_1, T_1) < x) > 0$, for all $x > 0$. Combined with the finite capacity and $\theta(x) < \infty$ for all finite x, this ensures that the workload process $\{D^{r(\cdot)}(t), t \in \mathbb{R}\}$ is still regenerative with customer arrivals into an empty system as regeneration points.

Define $\tilde{r}(x) := r(K - x)$, for $0 \leq x \leq K < \infty$, and let all random variables $X^{r(\cdot)}$ and $X^{\tilde{r}(\cdot)}$ correspond to the model with release rates $r(x)$ and $\tilde{r}(x)$, respectively, if the process is at level x. Similar to Section 2, we construct a "risk-type" 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 process is governed by the input rate function $\tilde{r}(x) = r(K - x)$ and satisfies

https://doi.org/10.1017/S0269964805050138 Published online by Cambridge University Press

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

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

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

Here, $\tilde{\theta}(x) := \int_0^x (r(y))^{-1} dy$, represents the time required to move from zero to x in the absence of negative jumps, with inverse $\tilde{\theta}^{-1}(t)$. Note that, for finite K, $\int_0^x (\tilde{r}(y))^{-1} dy < \infty$, meaning that state 0 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*

$$P_K^{r(\cdot)} = \mathbb{P}(C_{\max}^{\tilde{r}(\cdot)} \geq K) \tag{10}$$

or, alternatively,

$$P_K^{r(\cdot)} = \frac{\mathbb{P}(W^{K,r(\cdot)} = 0)}{\mathbb{P}(W^{K,\tilde{r}(\cdot)} = 0)} \mathbb{P}(C_{\max}^{r(\cdot)} \geq K). \tag{11}$$

We use a direct approach to show (10), thereby extending the proof in Section 2.1. To show (11), 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 (10).

PROOF OF (10): As noted earlier, 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 (3) and the construction of the process $\{R^{\tilde{r}(\cdot)}(t), t \in \mathbb{R}\}$, leads to

$$\mathbb{P}(C_{\max}^{r(\cdot)} \geq K) = \mathbb{P}(\exists n \leq \tau_0 : R^{\tilde{r}(\cdot)}(T_{n^-}) - S_n \leq 0). \tag{12}$$

The probability in (12) can be interpreted as the probability that state 0 is reached before $R^{\tilde{r}(\cdot)}(t)$ hits state K again, starting off from level K. Define $R_0 = K$ and $R_{n+1}^{\tilde{r}(\cdot)} = g(R_n^{\tilde{r}(\cdot)}, S_{n+1}, T_n)$, with

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

Following [7], we construct the dual function corresponding to the described process $\{R^{\tilde{r}(\cdot)}(t), t \in \mathbb{R}\}$, yielding

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

https://doi.org/10.1017/S0269964805050138 Published online by Cambridge University Press

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. We now complete the proof of (10) by combining the duality (4) between storage and risk processes with the expression (1) for P_K:

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

■

Next we turn to (11), which we show following the lines of Section 2.2.

Proof of (11): As mentioned earlier, 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 off from level K, the additional number of customer rejections is geometrically distributed with success parameter $1 - q_K := \mathbb{P}(\tau_K < \tau_0 | D^{r(\cdot)}(0) = K)$. Thus, we have to show that $q_K = \mathbb{P}(W^{K,\tilde{r}(\cdot)} > 0)$ and combine (5)–(8) to complete the proof.

We start with the construction of the "risk-type" 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, $R^{\tilde{r}(\cdot)}(t)$ hits level 0 again before it reaches level K. Interpreting our process as a monotone stochastic recursion with two absorbing barriers, we define $R_{n+1}^{\tilde{r}(\cdot)} = g(R_n^{\tilde{r}(\cdot)}, S_{n+1}, T_n)$, where

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

Again, using [7], 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 latter 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 does not equal the waiting time.

https://doi.org/10.1017/S0269964805050138 Published online by Cambridge University Press

Finally, using Corollary 3.1 of [7] once more, we obtain

$$\begin{aligned} q_K &= \lim_{\epsilon \downarrow 0} \mathbb{P}(R^{\tilde{r}(\cdot)}_{\gamma(\epsilon,K)} \leq 0) \\ &= \lim_{\epsilon \downarrow 0} \mathbb{P}(V^{\tilde{r}(\cdot)} \geq \epsilon) \\ &= \mathbb{P}(W^{\tilde{r}(\cdot)}_k > 0). \end{aligned} \tag{13}$$

Hence, by combining (5)–(8), and (13), we also have shown the second part of the result. ■

Remark 3.1: The constant $\mathbb{P}(W^{K,r(\cdot)} = 0)/\mathbb{P}(W^{K,\tilde{r}(\cdot)} = 0)$ in (10) can easily be interpreted. As the interarrival times in both systems follow the same distribution, using (6), the constant equals the ratio of the respective cycle lengths.

Remark 3.2: A sample path argument can also provide some intuition into the equivalence between (10) and (11). First, the process $\{R^{\tilde{r}(\cdot)}(t)|t \geq 0\}$ 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-type process into a queuing process again, we use a reversibility argument, as in Asmussen and Kella [4] and Asmussen and Schock Petersen [6]. The sample path of this queuing process can essentially be obtained by time-reversing the sample path of $\{R^{\tilde{r}(\cdot)}(t)|t \geq 0\}$, resulting in a queuing 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. Given the results derived earlier, this leads to more transparent proofs of existing results and to results that are new.

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 queuing models. We refer to Asmussen [2] for a survey of these results. The $M/G/1$ case has already been treated in [23]. 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 μ. Then*

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

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

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

https://doi.org/10.1017/S0269964805050138 Published online by Cambridge University Press

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

PROOF: The result follows immediately from Theorem 2.1 and formula (7.76) of Cohen [11], stating that for the $GI/M/1$ queue,

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

with $H(x), x \geq 0$, defined as earlier. ■

4.2. Asymptotics

Van Ommeren and De Kok [22] derived 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 $f(x)/g(x) \to 1$ as $x \to \infty$. Asymptotics for the latter are due to Iglehart [16]: Under certain regularity conditions (see [16]), it holds that

$$\mathbb{P}(C_{\max} \geq K) \sim D\mathrm{e}^{-\gamma K}, \tag{14}$$

for certain positive constants γ and D. Using Theorem 2.1, the proof of the main result of [19] is now trivial: Just combine Theorem 2.1 with (14) to (re-)obtain

$$P_K \sim D\mathrm{e}^{-\gamma K}.$$

For more details concerning specific assumptions and expressions for γ and D we refer to [16,22].

We conclude 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, Klüppelberg, and Mikosch [13] 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 [3]. 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) the following corollary.

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

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

Also, in case of Poisson arrivals, this result can be extended to queues with general service speeds; see [3] for details. Note that (10) and (11), combined with Remark 3.1, indeed result in the same asymptotics.

https://doi.org/10.1017/S0269964805050138 Published online by Cambridge University Press

4.3. Poisson Arrivals and Takács' Formula

The equivalence in Theorem 2.1 can also be used the other way around: Given information on P_K, we derive a new identity for the distribution of $C_{\max}$ for queues with a 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 assumptions of Poisson arrivals, and a stationary (embedded) workload distribution in case of infinite-buffer capacity (see, e.g., [1,9] for details). Under these assumptions, the steady-state distribution of the amount of work in the system found by a customer $W^{K,r(\cdot)}$ satisfies the following *proportionality* result:

$$\mathbb{P}(W^{K,r(\cdot)} \le x) = \frac{\mathbb{P}(W^{r(\cdot)} \le x)}{\mathbb{P}(W^{r(\cdot)} \le K)}. \tag{15}$$

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 [11], and Hooghiemstra [15]. For a rigorous proof of (15) in the case of a general release rate, we refer to Asmussen [1, Chap. XIV, Prop. 3.1].

Writing $1 - P_K = \mathbb{P}(W^{K,r(\cdot)} + S < K)$, conditioning on S, applying (15), and deconditioning on S then results in

$$\begin{aligned} P_K^{r(\cdot)} &= 1 - \mathbb{P}(W^{K,r(\cdot)} + S < K) \\ &= \frac{\mathbb{P}(W^{r(\cdot)} + S \ge K) - \mathbb{P}(W^{r(\cdot)} > K)}{\mathbb{P}(W^{r(\cdot)} \le K)}. \end{aligned}$$

Combining this result with (11) then results in the following corollary.

COROLLARY 4.3: *Assume that the $M/G/1$ queue with infinite buffer size and general release rate has a stationary (embedded) workload distribution. Then*

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

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 [10], and Asmussen and Perry [5] 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

$$\mathbb{P}(C_{\max} < x) = \frac{\mathbb{P}(W + S < x)}{\mathbb{P}(W \le x)}.$$

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 and Resnick [14] and Bekker [8]. 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

https://doi.org/10.1017/S0269964805050138 Published online by Cambridge University Press

to obtain asymptotics and/or bounds. Asymptotic results in the light-tailed case are hardly known; see Asmussen [2,3].

5. CONCLUSION

We have considered several queuing models that 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, we believe that an appropriate modification of Theorem 2.1 still holds for other queuing 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 and Perry [5].

Furthermore, we expect that Siegmund duality and related results can also be fruitful in other queuing problems. In the context of the present article, we believe that an analog of (2) can be shown for queues that can be modeled as birth–death processes: Siegmund-type duality results for birth and death processes have been derived by Dette, Fill, Pitman, and Studden [12].

Acknowledgments

This work was supported by a research grant from Philips Electronics (to R. B.) and a grant from NWO (to B. Z.).

References

1. Asmussen, S. (2003). *Applied probability and queues*, 2nd ed. New York: Springer-Verlag.
2. Asmussen, S. (1998). Extreme value theory for queues via cycle maxima. *Extremes* 2: 137–168.
3. Asmussen, S. (1998). Subexponential asymptotics for stochastic processes: Extremal behaviour, stationary distributions and first passage times. *Annals of Applied Probability* 8: 354–374.
4. Asmussen, S. & Kella, O. (1996). Rate modulation in dams and ruin problems. *Journal of Applied Probability* 33: 523–535.
5. Asmussen, S. & Perry, D. (1992). On cycle maxima, first passage problems and extreme value theory for queues. *Stochastic Models* 8: 421–458.
6. Asmussen, S. & Schock Petersen, S. (1988). Ruin probabilities expressed in terms of storage processes. *Advances in Applied Probability* 20: 913–916.
7. Asmussen, S. & Sigman, K. (1996). Monotone stochastic recursions and their duals. *Probability in the Engineering and Informational Sciences* 10: 1–20.
8. Bekker, R. (2004). Finite-buffer queues with workload-dependent service and arrival rates. SPOR Report 2004-01, Eindhoven University of Technology, The Netherlands.
9. Browne, S. & Sigman, K. (1992). Work-modulated queues with applications to storage processes. *Journal of Applied Probability* 29: 699–712.
10. Cohen, J.W. (1976). *Regenerative processes in queueing theory.* Berlin: Springer-Verlag.
11. Cohen, J.W. (1982). *The single server queue.* Amsterdam: North-Holland.
12. 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.
13. Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). *Modelling extremal events.* Berlin: Springer-Verlag.

14. 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.
15. Hooghiemstra, G. (1987). A path construction for the virtual waiting time of an $M/G/1$ queue. *Statistica Neerlandica* 41: 175–181.
16. Iglehart, D.G. (1972). Extreme values in the $GI/G/1$ queue. *Annals of Mathematical Statistics* 43: 627–635.
17. Lindley, D.V. (1959). Discussion of a paper by C.B. Winsten. *Proceedings of the Cambridge Philosophical Society* 48: 277–289.
18. 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.
19. Perry, D. & Stadje, W. (2003). Duality of dams via mountain processes. *Operations Research Letters* 31: 451–458.
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.* New York: Wiley.
22. 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.
23. Zwart, A.P. (2000). A fluid queue with a finite buffer and subexponential input. *Advances in Applied Probability* 32: 221–243.