Loss rates in the M/G/1 queue with complete rejection

TU/e technische universiteit eindhoven

SPOR-Report 2003-27

Loss rates in the *M/G/*1 queue with complete rejection

A.P. Zwart

SPOR-Report
Reports in Statistics, Probability and Operations Research

Eindhoven, November 2003
The Netherlands

/department of mathematics and computing science

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
Internet: http://www.win.tue.nl/math/bs/cosor.html

ISSN 1567-5211

Loss rates in the $M/G/1$ queue with complete rejection

Bert Zwart

Eindhoven University of Technology
Department of Mathematics & Computer Science
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
zwart@win.tue.nl

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

November 5, 2003

Abstract

Consider the $M/G/1$ queue in which customers are rejected if their total sojourn time would exceed a certain level K. A pasic performance measure of this system is the probability P_K that a customer gets rejected in steady state. This paper presents asymptotic expansions for P_K as $K \to \infty$. If the service time B is light-tailed, it is shown that the loss probability has an exponential tail. The proof of this result heavily relies on recent results on the two-sided exit problem for Lévy processes with no positive jumps. For heavy-tailed (subexponential) service times, the loss probability is shown to be asymptotically equivalent to the trivial lower bound $P(B > K)$.

2000 Mathematics Subject Classification: 60K25 (primary), 60J30, 68M20, 90B22 (secondary).

Keywords & Phrases: queues, storage processes, complete rejection, loss probability, Lévy processes, two-sided exit problem, asymptotic expansions, light tails, heavy tails.

1 Introduction

This paper considers the following variation of the $M/G/1$ queue: customers that arrive are accepted if and only if their total sojourn time is less than a fixed constant K. If this is not the case, then a customer is rejected completely. Thus the workload $W_{K,n}$ in the system before the n-th arrival is driven by the following recursion:

$$W_{K,n+1} = \begin{cases} (W_{K,n} + B_n - A_n)^+ & \text{if } W_{K,n} + B_n \leq K \\ (W_{K,n} - A_n)^+ & \text{if } W_{K,n} + B_n > K. \end{cases} \tag{1.1}$$

We are interested in the probability P_K that a customer is rejected in steady state, more precisely, in the behavior of P_K as $K \to \infty$. If the system load $\rho < 1$ (which we assume

throughout this paper) it is clear that $P_K \to 0$. This paper gives exact rates of convergence for both light-tailed and heavy-tailed service times.

The model decribed by (1.1) seems to have a special place in the literature on queueing models with rejection. In particular, it is not as well understood as the $M/G/1$ queue where customers are not completely but only *partially* rejected (i.e. part of a rejected customer's work is accepted such that the buffer is completely filled); this model is also known as the finite dam. The steady-state distribution of the workload in this queue is already known since Takács [21]. The probability P_K^p that a customer is (partially) rejected can be expressed in terms of the tail distribution of the maximum amount of work $V_{\max}$ in the system during a busy cycle of the infinite buffer queue. In particular, the following result (which even holds for the $GI/G/1$ queue with partial rejection) can be found and Bekker & Zwart [5]:

$$P_K^p = P(V_{\max} > K).$$

Another tractable model is the $M/G/1$ queue where customers leave the system due to impatience when their waiting time has exceeded a fixed treshold K. In this case, the probability of impatience P_K^i is equal to

$$P_K^i = \frac{(1-\rho)P(W_{M/G/1} > K)}{1 - \rho P(W_{M/G/1} > K)},$$

with $W_{M/G/1}$ the steady-state waiting time distribution in the $M/G/1$ queue, see Boots & Tijms [8]. These formulas can easily be applied to obtain asymptotic expansions for P_K^p or P_K^i, since the asymptotic behavior of $P(W_{M/G/1} > K)$ and $P(V_{\max} > K)$ is well known for both the light-tailed and the heavy-tailed case.

Unfortunately, such a simple program cannot be carried out for the $M/G/1$ queue with complete rejection. The main problem is the intractable distribution of the amount of work in the system when a customer is rejected. (In the case of partial rejection, this amount of work is always K.) Another problem with this queueing model is that its driving recursion (1.1) *fails to be monotone* in its main argument $W_{K,n}$. This rules out the possibility of relating P_K to a first passage probability using the framework of Asmussen and Sigman [1]. This approach has been proven quite fruitful when considering queues with partial rejection; see e.g. [5].

Nevertheless, special treatments are possible for the $M/M/1$ and $M/D/1$ queues; see Cohen [9], Gavish & Schweitzer [14] and Asmussen & Perry [2]. De Kok & Tijms [16] derived the asymptotic behavior of P_K in the $M/M/1$ case with service rate μ. In particular, they show that

$$P_{K,M/M/1} \sim (1-\rho)\mathrm{e}^{-\rho}\mathrm{e}^{-\mu(1-\rho)K}, \tag{1.2}$$

as $K \to \infty$, where $f(x) \sim g(x)$ means $\lim f(x)/g(x) = 1$. For the more general $M/G/1$ queue, it is conjectured in [16] that P_K has an exponential tail. This conjecture was only partially resolved by Van Ommeren [18], who obtained asymptotic lower and upper bounds.

The main goal of the present paper is to settle this conjecture for a general class of light-tailed service-times: it is shown that, for some constants D and γ,

$$P_K \sim De^{-\gamma K},$$

as $K \to \infty$. Unfortunately, the prefactor D in this expansion is quite difficult to compute. The expression we obtain for D is related to the solution of a certain Fredholm-type integral equation.
This result should be contrasted with the case where service times are heavy tailed (more precisely, when service times are in the class $\mathcal{S}^*$, see Section 2). In that case we show (even for the more general $GI/GI/1$ queue) that

$$P_K \sim P(B > K).$$

Thus the trivial lower bound $P_K \geq P(B > K)$ is attained as $K \to \infty$.
Not surprisingly, the methods we use to prove the asymptotic expansions for P_K strongly depend on whether service time are light-tailed or heavy-tailed. In the light-tailed case, we heavily rely on results on the two-sided exit problem for completely asymmetric Lévy processes (i.e. Lévy processes with no positive or no negative jumps). In present form, these results are known since Suprun [20], who approached the problem using Wiener-Hopf factorization. The results of [20] came available to a wider audience in Bertoin [7]. The latter paper attacks the two-sided exit problem using excursion theory. A recent survey containing martingale proofs is Kyprianou [17]. The results which are of direct use for us are collected in Section 4. Using these results, we are able to obtain an expression for the distribution of the amount of work *right before* a loss occurs. This distribution provides the key to deriving the asymptotics. When service times are heavy-tailed, the key is to show is that the system workload is $O(1)$ (as the buffer size $K \to \infty$) when a customer is rejected. This is possible by exploiting some estimates due to Asmussen [3] and Foss & Zachary [13].
This paper is organized as follows: a detailed model description of the $M/G/1$ queue with complete rejection, as well as some auxiliary results on the $M/G/1$ queue with infinite buffer size, are given in Section 2. We present our main results in Section 3. Section 4 is devoted to the two-sided exit problem for Lévy processes with no positive jumps. These results are then applied in Section 5 to obtain a proof of the asymptotics for P_K in the light-tailed case. A proof of the heavy-tailed asymptotics can be found in Section 6.

2 Preliminaries

This section contains several preliminary results. We start with a description of the workload process of the $M/G/1$ queue with complete rejection. Then we give several asymptotic results for the single server queue without rejection which are used in this paper.

2.1 The $M/G/1$ queue with complete rejection

Customers arrive according to a Poisson process with rate λ. Service times are given by the i.i.d. sequence $B_i, i \geq 1$. A generic service time is denoted by B, and has Laplace-Stieltjes transform (LST) $\beta(s)$. Throughout the paper, it is assumed that $\rho = \lambda E[B] < 1$.
The workload process in the $M/G/1$ queue with complete rejection is defined as follows: Let $T_1, T_2, \ldots$ 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., $\bar{T}_n = \sum_{k=1}^{n} T_k$. The workload process $\{V_K(t), t \in \mathbb{R}\}$ is then defined recursively by

$$V_K(t) = \max(V_K(\bar{T}_k^-) + B_k I_{(V_K(\bar{T}_k^-)+B_k \leq K)} - (t - \bar{T}_k), 0), \qquad t \in [\bar{T}_k, \bar{T}_{k+1}), \quad (2.3)$$

where $I_{(\cdot)}$ is the indicator function. The workload process $\{V_K(t), t \in \mathbb{R}\}$ is regenerative, with customer arrivals into an empty system being regeneration points.

2.2 The single-server queue with infinite buffer size

Our analysis partly relies on several results for the standard single server queue. In particular, we need the tail behavior of the waiting-time distribution, and the tail behavior of the distribution of the maximum workload during a busy cycle; these results are gathered in this section.
As mentioned in the introduction, we both consider light-tailed and heavy-tailed asymptotics. When we assume that the service time distribution is light tailed, we mean the following:

Assumption L
There exists a constant $\gamma > 0$ such that

$$\frac{\lambda}{\lambda + \gamma} E[e^{\gamma B}] = 1, \quad (2.4)$$

$$E[B e^{\gamma B}] < \infty. \quad (2.5)$$

If Assumption L is valid, then the tail of the waiting-time distribution in the $M/G/1$ queue satisfies:

$$P(W_{M/G/1} > u) \sim C e^{-\gamma u}, \qquad u \to \infty. \quad (2.6)$$

The constant C is given by $C = (1 - \rho)/(\lambda E[B e^{\gamma B}] - 1)$. This result, due to Lundberg, is classical and can be found in most applied probability textbooks; see for example Theorem XIII.5.2 of Asmussen [4].

A similar result holds for the maximum amount of work during a cycle, defined as $V_{\max}$. The following result is due to Iglehart [15], and is again valid under Assumption L:

$$P(V_{\max} > u) \sim C_0 e^{-\gamma u}, \quad (2.7)$$

with $C_0 = C(E[e^{\gamma B}] - 1)$, where C is the same constant which appears in (2.6).

The above results are all concerned with light-tailed service times. In this paper we call service times heavy-tailed if they belong to the class $\mathcal{S}^*$, i.e.

Assumption H
Let $F(x) = P(B \leq x)$ and $\bar{F}(x) = 1 - F(x)$. Then,

$$\lim_{x\to\infty} \int_0^x \frac{\bar{F}(x-y)}{\bar{F}(x)} \bar{F}(y)\mathrm{d}y = 2E[B].$$

If Assumption H holds, then the following asymptotic estimate holds, even for the $GI/GI/1$ queue; see Asmussen [3] and Foss & Zachary [13]:

$$P(V_{\max} > K) \sim E[N]P(B > K). \tag{2.8}$$

The prefactor $E[N]$ is the expected number of customers arriving during one busy cycle. Foss & Zachary [13] also show a converse result: if (2.8) holds, then the service time distribution satisfies Assumption H. For background on heavy tails, we refer to the monograph Embrechts *et al.* [12].

3 Main results

In this section we present the main results of this paper, i.e. asymptotic expansions for P_K under light-tailed and heavy-tailed assumptions. We first present our result for light-tailed service times. Define

$$\begin{aligned}
W(x) &= P(W_{M/G/1} \leq x)/(1-\rho), && (3.9)\\
Q(x,y) &= [W(x) - I_{(x\geq y)}W(x-y)]\lambda P(B > y), && (3.10)\\
Q_1(x,y) &= Q(x,y),\\
Q_n(x,y) &= \int_{z=0}^{\infty} Q_{n-1}(x,z)Q(z,y)\mathrm{d}z, \quad n \geq 2,\\
Q^*(x,y) &= \sum_{n=1}^{\infty} Q_n(x,y).
\end{aligned}$$

With these definitions we are able to state our first theorem:

Theorem 3.1. *Assume that the arrival process is Poisson, let $\rho < 1$, and assume that the service-time distribution satisfies Assumption L. Then there exists a constant $D \in (0,\infty)$ such that*

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

The prefactor D can be written as

$$D = (1-\rho)C_0 D_0,$$

with C_0 given below (2.7) and

$$D_0 = 1 + \int_{y=0}^{\infty}\int_{x=0}^{\infty} Q^*(x,y)\frac{\mathrm{e}^{\gamma x}-1}{1-\rho}\lambda P(B > x)\mathrm{d}x\mathrm{d}y. \tag{3.11}$$

Thus, as conjectured in De Kok & Tijms [16], the probability P_K indeed has an exponential tail. Unfortunately, the prefactor D is very difficult to compute; especially when using the expression given above. Recall that for the $M/M/1$ queue, D can be computed: it is shown in De Kok & Tijms [16] that $D = (1-\rho)e^{-\rho}$, cf. (1.2). Note that $Q^*(x,y)$ can be viewed as the solution of a Fredholm-type integral equation with kernel $Q(x,y)$. The relation between such equations and queues with rejection has been observed before in [2]. A probabilistic interpretation of $Q(x,y)$ is given in Section 4.

As the next result shows, the asymptotics for P_K in the heavy-tailed case are much easier to describe. Moreover, it is not necessary to consider Poisson arrivals:

Theorem 3.2. *Assume that the arrival process is a renewal process, let $\rho < 1$, and assume that the service-time distribution satisfies Assumption H. Then*

$$P_K \sim P(B > K).$$

Thus, the trivial lower bound $P_K \geq P(B > K)$ is asymptotically exact when service times are heavy tailed. Theorem 3.2 reveals that, in the heavy-tailed case, a customer is most likely rejected since its own service time is large. Right before (thus also right *after*) rejection, the workload in the system is O(1) as $K \to \infty$.
In the proof of Theorems 3.1 and 3.2 we use the following representation for P_K. Let N_K denote the number of customers arriving during a busy period, and let L_K the number of customers lost during a busy cycle. Then, using the theory of regenerative processes, we obtain

$$\begin{aligned} P_K &= \frac{E[L_K]}{E[N_K]} \\ &= \frac{E[L_K \mid L_K \geq 1]}{E[N_K]} P(L_K \geq 1) \\ &= \frac{E[L_K \mid L_K \geq 1]}{E[N_K]} P(V_{\max} \geq K). \end{aligned}$$

In the third equality, we used the obvious identity $P(L_K \geq 1) = P(V_{\max} \geq K)$.
With this representation at our disposal, the idea of the proof is clear: In both the light-tailed and the heavy-tailed case, it holds that $E[N_K] \to E[N]$ (which equals $1/(1-\rho)$ in the $M/G/1$ queue). Furthermore, the asymptotic behavior of $P(V_{\max} \geq K)$ is given in Subsection 2.2, both under Assumption L and Assumption H. Thus, it remains to show that $E(L_K \mid L_K \geq 1)$ converges to a constant as $K \to \infty$. In Section 6 we show that this constant converges to 1 if service-times are heavy-tailed. Obtaining the limit of $E(L_K \mid L_K \geq 1)$ under light-tailed assumptions (which equals D_0) is much more involved. This requires several non-trivial results on Lévy processes which are given in the following section.

4 The two-sided exit problem

This section concentrates on the two-sided exit problem and paves the way to the proof of Theorem 3.1, which is the subject of the next section. We use the same notation as

Bertoin [7]: consider a Lévy process $X_t, t \geq 0$ with no positive jumps. Define $P_x(\cdot)$ as $P(\cdot \mid X_0 = x)$, and set $P = P_0$. The distribution of X_t is given by its moment generating function

$$E(\mathrm{e}^{sX_t}) = \mathrm{e}^{t\psi(s)}.$$

An important special case (in view of our queueing application) is when

$$X_t = t - \sum_{i=1}^{N_t} B_i, \tag{4.12}$$

with (as in the previous sections) $B_i, i \geq 1$ an i.i.d. sequence with common LST $\beta(s)$, and $N_t, t \geq 0$ a Poisson process with rate λ. In that case,

$$\psi(s) = s - \lambda(1 - \beta(s)).$$

Fix a, and define

$$T = \inf\{t : X_i \notin (0, a)\}.$$

Let Δ_T be the jump at time T, i.e., $\Delta_T = X_T - X_{T-}$. This section presents the joint distribution of X_{T-} and Δ_T, both for fixed a and $a \to \infty$.
First, we treat the case of fixed a. We start with a classical result (Takács [21]):

$$P_x(X_T = a) = W(x)/W(a), \tag{4.13}$$

with $W : [0\,\infty) \to [0, \infty)$ the unique continuous function such that

$$\int_0^\infty \mathrm{e}^{-sx} W(x)\mathrm{d}x = \frac{1}{\psi(s)}.$$

The function W is known as the *scale function*; if X_t is compound Poisson, one can relate W to the steady state waiting time distribution in the $M/G/1$ queue if the latter exists, cf. (3.1). The joint distribution of X_{T-} and Δ_T has been given in Bertoin [7]; see also Suprun [20]. In the present paper, we only need Corollary 2 of [7], which is restated in the following proposition.

Proposition 4.1. *(Bertoin [7]) For every $x, y \in (0, a)$ and every $z \leq -y$ we have*

$$P_x(X_{T-} \in dy; \Delta_T \in dz) = \left(\frac{W(x)W(a-y)}{W(a)} - I_{(x \geq y)} W(x-y)\right) \Lambda(dz)$$

where Λ denotes the Lévy measure of X. In particular,

$$Q(a, x, y) := P_x(X_{T-} \in dy; X_T \leq 0) = \left(\frac{W(x)W(a-y)}{W(a)} - I_{(x \geq y)} W(x-y)\right) \Lambda(y, \infty). \tag{4.14}$$

Using this proposition, we now derive the asymptotic distribution of (X_{T-}, Λ_T) under the assumption that X_t is of the form (4.12) and that X_t has a positive drift. Under (4.12), the latter assumption is equivalent to

$$E(X(1)) = 1 - \lambda E(B) = 1 - \rho > 0.$$

Note that, when (4.12) holds, the Lévy measure in Proposition 4.1 is given by

$$\Lambda(dz) = \lambda \mathrm{d}\mathbb{P}(B \leq z).$$

Using Proposition 4.1 we obtain the following result.

Proposition 4.2. *Assume that X_t is compound Poisson as in (4.12) with $\rho < 1$ and that Assumption L holds. Then, as $a \to \infty$, for each x,*

$$P_{a-x}(X_{T-} \in \mathrm{d}y; \Lambda_T \in -\mathrm{d}z \mid X_T \leq 0) \to \frac{\mathrm{e}^{\gamma y} - 1}{1 - \rho} \lambda \mathrm{d}P(B \leq z). \tag{4.15}$$

In particular,

$$P_{a-x}(X_{T-} \in \mathrm{d}y \mid X_T \leq 0) \to \frac{\mathrm{e}^{\gamma y} - 1}{1 - \rho} \lambda P(B > y). \tag{4.16}$$

This proposition gives the asymptotic distribution of the level of X_t *right before* jumping below 0. As one can see, the asymptotic distribution is independent of the level x, which is not very surprising.

Proof. The proof follows from direct computations. Fix x, y, z and write for $a > x + y$, using Proposition 4.1 and (4.13),

$$P_{a-x}(X_{T-} \in \mathrm{d}y; \Lambda_T \in -\mathrm{d}z \mid X_T \leq 0) = \frac{W(a-x)W(a-y) - W(a)W(a-x-y)}{W(a) - W(a-x)} \lambda \mathrm{d}P(B \leq z).$$

We treat the numerator and denominator on the right hand side of this expression separately. First, we analyze the denominator. Using (2.6), it follows that, as $a \to \infty$,

$$W(a) = \frac{1}{1-\rho} - \frac{C}{1-\rho} \mathrm{e}^{-\gamma a}(1 + \mathrm{o}(1)). \tag{4.17}$$

This implies

$$W(a) - W(a-x) \sim C \frac{\mathrm{e}^{\gamma x} - 1}{1 - \rho}.$$

To obtain the asymptotic behavior of the numerator, we apply (4.17) four times. A simple computation then gives

$$W(a-x)W(a-y) - W(a)W(a-x-y) \sim \frac{C\mathrm{e}^{-\gamma a}(1 + \mathrm{o}(1))}{(1-\rho)^2} \left[1 + \mathrm{e}^{\gamma(x+y)} - \mathrm{e}^{\gamma x} - \mathrm{e}^{\gamma y}\right].$$

This implies

$$\begin{aligned} \frac{W(a-x)W(a-y) - W(a)W(a-x-y))}{W(a) - W(a-x)} &\to \frac{1}{1-\rho} \frac{1}{\mathrm{e}^{\gamma x} - 1} \left[\mathrm{e}^{\gamma y}(\mathrm{e}^{\gamma x} - 1) - (\mathrm{e}^{\gamma x} - 1)\right] \\ &= \frac{\mathrm{e}^{\gamma y} - 1}{1 - \rho}, \end{aligned}$$

which completes the proof. □

The previous result provided the asymptotic distribution when one starts at a high level $a - x$, i.e. close to a. We also need the asymptotic distribution as $a \to \infty$ when we start at level x (i.e., close to 0); this is presented in the next proposition.
Recall that $Q(a, x, y) = P_x(X_{T-} \in \mathrm{d}y; X_T \leq 0)$.

Proposition 4.3. *As* $a \to \infty$,

$$Q(a, x, y) \to Q(x, y) = [W(x) - I_{(x \geq y)} W(x - y)] \lambda P(B \geq y).$$

Proof. A straightforward combination of Proposition 4.1 and (4.17). □

We close this section with some remarks:

- The function $Q(x, y)$, appearing as limit in Proposition 4.3 and already defined in Section 3, can be interpreted as as follows: consider a risk process with initial capital x. Then $Q(x, y)\mathrm{d}y$ is the probability that ruin eventually occurs, and that the surplus before ruin is in the interval $(y, y + \mathrm{d}y)$. The distribution of the surplus prior to ruin has been in investigated in Schmidli [19].

- Both Proposition 4.2 and 4.3 are for compound Poisson processes. This assumption can be relaxed: asymptotics for the scale function $W(x)$ without the assumption (4.12) can be derived from results in Bertoin & Doney [6], who prove an analogue of (2.6) for the supremum of a Lévy process. Since our primary interest is in the compound Poisson case, we omit the details.

We now turn to an analysis of the loss probability P_K.

5 Proof of Theorem 3.1

In this section we give a proof of Theorem 3.1, which states the asymptotics for P_K under the (light tail) Assumption L. Recall that

$$P_K = \frac{E[L_K \mid L_K \geq 1]}{E[N_K]} P(V_{\max} \geq K).$$

By monotone convergence we have $E[N_K] \to E[N] = 1/(1 - \rho)$, and from (2.7) we obtain $P(V_{\max} \geq K) \sim C_0 \mathrm{e}^{-\gamma K}$. Thus, to prove Theorem 3.1, it suffices to show that, under Assumption L and $\rho < 1$,

$$E[L_K \mid L_K \geq 1] \to D_0, \tag{5.18}$$

with D_0 defined as in Section 3. Write

$$E[L_K \mid L_K \geq 1] = \sum_{n=1}^{\infty} P(L_K \geq n \mid L_K \geq 1)$$

We now obtain an expression for $P(L_K \geq n \mid L_K \geq 1)$ in terms of the undershoot probabilities $Q(a, x, y)$, as derived in the previous section. For this, it will be convenient

to work with the process $R_K(t) = K - V_K(t)$ representing the spare capacity of the buffer at time t; recall that $V_K(t)$ is the workload at time t as defined in Section 2.1. Let t_n be the time of the n-th rejection in a cycle. We take $t_n = \infty$ if $L_K < n$. Define for $n \geq 2$ the following densities:

$$p_{K,n}(x,y) = P(L_K \geq n; R_K(t_n) \in \mathrm{d}y \mid R_K(t_{n-1}) = x; L_K \geq n-1). \tag{5.19}$$

Then, using the strong Markov property, it is obvious that for $n \geq 2$,

$$p_{K,n}(x,y) = Q(K,x,y). \tag{5.20}$$

Set

$$p_{K,n}(y) = P(L_K \geq n; R_K(t_n) \in \mathrm{d}y \mid L_K \geq 1). \tag{5.21}$$

Then, for $n \geq 2$,

$$\begin{aligned} p_{K,n}(y) &= \int_{0+}^{K} p_{K,n}(x,y) p_{K,n-1}(x) \mathrm{d}x \\ &= \int_{0+}^{K} Q(K,x,y) p_{K,n-1}(x) \mathrm{d}x. \end{aligned}$$

It remains to specify $p_{K,1}(x)$. This probability is given by

$$p_{K,1}(y) = \int_{0}^{K} Q(K, K-u, y) \mathrm{d}P(B \leq u). \tag{5.22}$$

Finally, note that for $n \geq 2$,

$$P(L_K \geq n \mid L_K \geq 1) = \int_{0}^{K} p_{K,n}(y) \mathrm{d}y. \tag{5.23}$$

We now let $K \to \infty$. Then, using Proposition 4.3 and (5.20), we obtain

$$p_{K,n}(x,y) \to Q(x,y). \tag{5.24}$$

We now inductively prove that the quantities $p_{K,n}(x)$ converge. We start with $n = 1$. Using Proposition 4.2 we obtain

$$Q(K, K-u, y) \to \frac{\mathrm{e}^{\gamma y} - 1}{1-\rho} \lambda P(B > y) =: p_1(y). \tag{5.25}$$

It is not difficult to show that for each y, $Q(K,x,y)$ is bounded in K and $x, 0 \leq x \leq K$. Thus, using the bounded convergence theorem, we obtain

$$p_{K,1}(y) \to p_1(y). \tag{5.26}$$

From this, we readily obtain by an inductive argument:

$$p_{K,n}(y) \to p_n(y) = \int_{0+}^{\infty} Q(x,y) p_{n-1}(x) \mathrm{d}x. \tag{5.27}$$

Finally, we obtain that, for $n \geq 2$,

$$P(L_K \geq n \mid L \geq 1) \to p_n := \int_0^\infty p_n(y)\mathrm{d}y. \tag{5.28}$$

Thus, since $p_1 = 1$, we conclude that

$$E[L_K \mid L_K \geq 1] \to \sum_{n=1}^{\infty} p_n = \sum_{n=1}^{\infty} \int_0^\infty p_n(y)\mathrm{d}y. \tag{5.29}$$

That this quantity equals D_0 as given by (3.11) can easily be verified by iterating (5.27). This completes the proof of Theorem 3.1.

6 Proof of Theorem 3.2

In this Section, it is assumed that Assumption H is in force. Starting point is again the expression

$$P_K = \frac{1}{E[N_K]} E[L_K \mid L_K \geq 1] P(C_{\max} > K).$$

Since, cf. (2.8),

$$P(V_{\max} > K) \sim E[N] P(B > K),$$

and since $E[N_K] \to E[N]$, it suffices to show that

$$E[L_K \mid L_K \geq 1] \to 1. \tag{6.30}$$

To prove this, we use an estimate due to Foss & Zachary [13]. Since B is in particular long-tailed, there exists a function $h(x) = o(x)$ with $h(x) \to \infty$ as $x \to \infty$ such that $P(B > x) \sim P(B > x - h(x))$. Recall that t_1 is the first time a customer gets rejected. We now have the following fact [13]:

$$P(V_K(t_1-) > h(K) \mid L_K \geq 1) \to 0.$$

Now write

$$\begin{aligned} E[L_K \mid L_K \geq 1] &= E[L_K 1_{(V_K(t_1-) \leq h(K)} \mid L_K \geq 1] \\ &+ E[L_K 1_{(V_K(t_1-) > h(K)} \mid L_K \geq 1] \\ &= I + II. \end{aligned}$$

We first prove that term I converges to 1 and then show that $II \to 0$. In both cases it suffices to prove the upper bound, the lower bound being trivial. To achieve an upper bound, we assume that the service discipline is changed into *partial* rejection *after* time t_1. This gives a sample-path wise increase of the workload process; thus it does not decrease the number of losses until the system empties. Denote the number of losses in the partial rejection model by L_K^p. It is shown in [5] that $L_K^p \mid L_K^p \geq 1$ has a geometric distribution

with rate $1/E[N_K]$. This implies that $E[L_K^p \mid L_K^p \geq 1] = E[N_k] \leq E[N]$. We shall use these results below.

Term I
As a worst case, we take $V_K(t_1) = V_K(t_1-) = h(K)$. It is clear that the probablity of a loss after time t_1 and before the queue empties is $o(1)$ as $K \to \infty$. Given that this occurs, the number of losses after time t_1 is geometrically distributed with rate $1/E[N_K]$. Thus the expected number of losses, given that a loss occurs, equals $E[N_K] \leq E[N]$. From this, we conclude that

$$I \leq 1 + E[N]o(1).$$

Term II
Assume now, to obtain an upper bound, that the system starts at level K at time t_1. The number of additional customers that get rejected is again geometrically distributed with rate $1/E[N_K]$. Thus, as $K \to \infty$,

$$II \leq E[N]P(V_K(t_1-) > h(K) \mid L_K \geq 1) \to 0.$$

This concludes the proof of Theorem 3.2.

Acknowledgments
The author is grateful to Henk Tijms for posing the problem and to René Bekker for comments on an earlier draft of this paper.

References

[1] Asmussen, S., Sigman, K. (1996). Monotone stochastic recursions and their duals. *Probability in the Engineering and Informational Sciences* **10**, 1–20.

[2] Asmussen, S., Perry, D. (1996). Rejection rules in the $M/G/1$ queue. *Queueing Systems* **19**, 105–130.

[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. (2003). *Applied Probability and Queues.* Second edition. Springer, New York.

[5] Bekker, R., Zwart, A.P. (2003). On an equivalence between loss rates and cycle maxima in queues and dams. SPOR Report 2003-17, Department of Mathematics and Computer Science, Eindhoven University of Technology. Submitted for publication.

[6] Bertoin, J. Doney, R. (1994). Cramér's estimate for Lévy processes. *Statistics and Probability Letters* **21**, 363–365.

[7] Bertoin, J. (1997). Exponential Decay and Ergodicity of completely asymmetric Lévy processes on a finite interval. *Annals of Applied probability* **7**, 156–169.

[8] Boots, N.K., Tijms, H.C. (1999). A multiserver queueing system with impatient customers. *Management Science* **45**, 444-448.

[9] Cohen, J.W. (1969). Single-server queues with restricted accessibility. *Journal of Engineering Mathematics* **3**, 265–284.

[10] Cohen, J.W. (1976). *Regenerative Processes in Queueing Theory.* Springer, Berlin.

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

[12] Embrechts, P., Klüppelberg, C., Mikosch, T. (1997). *Modelling Extremal Events.* Springer, Berlin.

[13] Foss, S. Zachary, S. (2003). The maximum on a random time interval of a random walk with long-tailed increments and negative drift. *Annals of Appliec Probability* **13**, 37–53.

[14] Gavish, B., Schweitzer, P. (1977). The Markovian Queue with bounded waiting time. *Management Science* **23**, 1349–1357.

[15] Iglehart, D.G. (1972). Extreme values in the $GI/G/1$ queue. *Annals of Mathematical Statistics* **43**, 627–635.

[16] de Kok, A.G., Tijms, H.C. (1985). A two-moment approximation for a buffer design problem requiring a small rejection probability. *Performance Evaluation* **5**, 77–84.

[17] Kyprianou, A.E. (2003). A martingale review of some fluctuation theory for spectrally negative Lévy processes. University of Utrecht, submitted for publication.

[18] van Ommeren, J.C.W. (1987). Exponential bounds for excess probabilities in systems with a finite capacity. *Stochastic processes and their applications* **24**, 143–149.

[19] Schmidli, H. (1999). On the distribution of the surplus prior and at ruin. *ASTIN Bulletin* **29**, 227–244.

[20] Suprun, V.N. (1976). Problem of destruction and resolvent of terminating processes with independent increments. *Ukranian mathematical journal* **28**, 39–45.

[21] Takács, L. (1967). *Combinatorial Methods in the Theory of Stochastic Processes.* Wiley, New York.

[22] Tijms, H.C. (2003). *A first course in stochastic models.* Wiley, New York.

[23] Zwart, A.P. (2000). A fluid queue with a finite buffer and subexponential input. *Advances in Applied Probability* **32**, 221–243.