Math Meth Oper Res (2015) 81:299–315
DOI 10.1007/s00186-015-0497-x

ORIGINAL ARTICLE

Loss rates in the single-server queue with complete rejection

Bert Zwart

Received: 14 March 2013 / Published online: 14 March 2015

Abstract Consider the single-server queue in which customers are rejected if their total sojourn time would exceed a certain level K. A basic 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 and inter-arrival times are exponential, it is shown that the loss probability has an exponential tail. The proof of this result heavily relies on results on the two-sided exit problem for Lévy processes with no positive jumps. For heavy-tailed (subexponential) service times and generally distributed inter-arrival times, the loss probability is shown to be asymptotically equivalent to the trivial lower bound $P(B > K)$.

Keywords Queues · Complete rejection · Loss probability · Lévy processes · Two-sided exit problem · Asymptotic expansions

Mathematics Subject Classification 60K25 (primary) · 60J30 · 68M20 · 90B22 (secondary)

1 Introduction

This paper considers the following variation of the single-server 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:

B. Zwart (✉)
CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
e-mail: bert.zwart@cwi.nl

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

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 described 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 single-server 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 (1967). 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 holds for the $GI/G/1$ queue with partial rejection) can be found in Bekker and Zwart (2005):

$$P_K^p = P(V_{\max} > K). \tag{1.2}$$

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 threshold K. In this case, the probability of abandonment P_K^i is equal to

$$P_K^i = \frac{(1-\rho)P(W_{M/G/1} > K)}{1 - \rho P(W_{M/G/1} > K)}, \tag{1.3}$$

with $W_{M/G/1}$ the steady-state waiting time distribution in the $M/G/1$ queue, see Boots and Tijms (1999). 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 single-server queue with complete rejection, even when we restrict ourselves to the $M/G/1$ case. 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 (1996). This approach has been proven quite fruitful when considering queues with partial rejection; see e.g. Bekker and Zwart (2005).

Nevertheless, special treatments are possible for the $M/M/1$ and $M/D/1$ queues; see Cohen (1969), Gavish and Schweitzer (1977) and Asmussen and Perry (1995). De Kok and Tijms (1985) derived the asymptotic behavior of P_K in the $M/M/1$ case with service rate μ. In particular, they show that

$$P_{K,M/M/1} \sim (1-\rho)\mathrm{e}^{-\rho}\mathrm{e}^{-\mu(1-\rho)K}, \tag{1.4}$$

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 Kok and Tijms (1985) that P_K has an exponential tail. This conjecture was only partially resolved by van Ommeren (1987), 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,

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

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 Sect. 2). In that case we show (for the $GI/GI/1$ 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 times 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); this is also the main reason that we restrict to exponentially distributed inter-arrival times. In present form, these results are known since Suprun (1976), who approached the problem using Wiener-Hopf factorization. The results of Suprun (1976) came available to a wider audience in Bertoin (1997). The latter paper attacks the two-sided exit problem using excursion theory. A survey containing martingale proofs is Kyprianou and Palmowski (2004). More recent papers on Lévy processes, queues, and partial rejection mechanisms are Asmussen and Pihlsgaard (2007); Debicki and Mandjes (2012); Norvang Andersen (2011)

The results which are of direct use for us are collected in Sect. 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 main point is to show 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 (1998) and Foss and Zachary (2003). The methods we use here allow for generally distributed inter-arrival times.

This paper is organized as follows: a detailed model description of the single-server queue with complete rejection, as well as some auxiliary results on the single-server queue with infinite buffer size, are given in Sect. 2. We present our main results in Sect. 3. Section 4 is devoted to the two-sided exit problem for Lévy processes with no positive jumps. These results are then applied in Sect. 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 Sect. 6. Section 7 puts our results in perspective by comparing them with other rejection disciplines.

2 Preliminaries

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

2.1 The single-server queue with complete rejection

We develop a description of the workload process in the single-server queue with complete rejection. Let $T_1, T_2, \ldots$ be the inter-arrival times of the customers and denote the arrival epoch of the n-th customer after time 0 by $\bar{T}_n$, i.e., $\bar{T}_n = \sum_{k=1}^{n} T_k$. 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 $\{V_K(t), t \in \mathbb{R}\}$ is then recursively defined by

$$V_K(t) = \max(V_K(\bar{T}_k^-) + B_k I_{(V_K(\bar{T}_k^-)+B_k \leq K)} - (t - \bar{T}_k), 0), \qquad t \in [\bar{T}_k, \bar{T}_{k+1}), \tag{2.1}$$

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. Assume that $V_K(0) = W_0^K = 0$ and that a job arrives at time 0. Let $C_K = \inf\{t > 0 : V_K(t) = 0\}$ be the length of a busy period and N_K be the number of customers served in that period. We drop the constant K from the notation if we consider the system with $K = \infty$.

Whenever service times are light-tailed (we make this more precise later on), we also assume that customers arrive according to a Poisson process with rate λ.

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 consider both 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

$$\frac{\lambda}{\lambda + \gamma} E[\mathrm{e}^{\gamma B}] = 1, \tag{2.2}$$

$$E[B\mathrm{e}^{\gamma B}] < \infty. \tag{2.3}$$

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

$$P(W_{M/G/1} > u) \sim Ce^{-\gamma u}, \quad u \to \infty. \tag{2.4}$$

The constant C is given by $C = (1-\rho)/(\lambda E[Be^{\gamma B}]-1)$. 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 (2003).

A similar result holds for the maximum amount of work during a cycle, defined as $V_{\max} = \max_{t\in[0,C]} V(t)$. The following result is due to Iglehart (1972), and is again valid under Assumption L:

$$P(V_{\max} > u) \sim C_0 e^{-\gamma u}, \tag{2.5}$$

with $C_0 = C(E[e^{\gamma B}] - 1) = C\gamma/\lambda$, where C is the same constant which appears in (2.4).

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 $F(x) = P(B \le x)$ 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 for the maximum waiting time during a busy cycle $W_{\max}$, even for the $GI/GI/1$ queue; see Asmussen (1998) and Foss and Zachary (2003):

$$P(W_{\max} > K) \sim E[N]P(B > K). \tag{2.6}$$

Foss and Zachary (2003) also show a converse result: if (2.7) holds, then the service time distribution satisfies Assumption H. For background on heavy tails, we refer to the monograph Embrechts et al. (1997). Note that $V_{\max} \ge W_{\max}$, and that $V_{\max} \le W_{\max} + a$ if inter-arrival times are a.s. bounded by a finite constant a. We will use this fact in Sect. 6; it can in fact be used to show the following result, which we could not find explicitly recorded in the literature:

Proposition 2.1 *For the $GI/GI/1$ queue, if B satisfies Assumption H,*

$$P(V_{\max} > K) \sim E[N]P(B > K). \tag{2.7}$$

Proof That $\liminf_{K\to\infty} P(V_{\max} > K)/P(B > K) \ge E[N]$ follows from (2.6) and the bound $V_{\max} \ge W_{\max}$. To achieve an upper bound, make the inter-arrival times smaller by truncating them at a finite constant a, and let N^a be the resulting number of customers in a busy period; note that $N \le N^a$. Since $V_{\max} \le W_{\max} + a$, since by Assumption H $P(B > K) \sim P(B > K - a)$, and by (2.6), we see that

$$\limsup_{K\to\infty} \frac{P(V_{\max} > K)}{P(B > K)} \le \limsup_{K\to\infty} \frac{P(W_{\max} > K - a)}{P(B > K)} = E[N^a].$$

The proof is now completed by noting that, due to bounded convergence, $E[N^a] \to E[N]$ as $a \to \infty$. □

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

$$\begin{aligned} W(x) &= P(W_{M/G/1} \leq x)/(1-\rho), \qquad (3.1)\\ q(x,y) &= [W(x) - I_{(x \geq y)} W(x-y)]\lambda P(B > y),\\ q_1(x,y) &= q(x,y),\\ q_n(x,y) &= \int_{z=0}^{\infty} q_{n-1}(x,z) q(z,y) dz, \qquad n \geq 2,\\ Q^*(x,y) &= \sum_{n=1}^{\infty} q_n(x,y). \qquad (3.2) \end{aligned}$$

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 as in (2.5) and

$$D_0 = 1 + \int_{y=0}^{\infty} \int_{x=0}^{\infty} Q^*(x,y) \frac{\mathrm{e}^{\gamma x} - 1}{1-\rho} \lambda P(B > x) dx dy. \qquad (3.3)$$

Thus, as conjectured in De Kok and Tijms (1985), 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 and Tijms (1985) that $D = (1-\rho)\mathrm{e}^{-\rho}$, 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 Asmussen and Perry (1995). A probabilistic interpretation of $q(x,y)$ is given in Sect. 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 Subsect. 2.2, both under Assumption L and Assumption H. Thus, it remains to be shown that $E(L_K \mid L_K \geq 1)$ converges to a constant as $K \to \infty$. In Sect. 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 (1997): 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

$$X_t = t - \sum_{i=1}^{N_t} B_i, \tag{4.1}$$

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

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

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 1967):

$$P_x(X_T = a) = W(x)/W(a), \tag{4.2}$$

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

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

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 (1997); see also Suprun (1976). In the present paper, we only need Corollary 2 of Bertoin (1997), which is restated in the following proposition.

Proposition 4.1 (Bertoin 1997) *For every $x, y \in (0, a)$ and every $z \leq -y$ we have*

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

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

$$\begin{aligned} Q(a, x, y) &:= P_x(X_{T-} \in dy; X_T \leq 0) \\ &= \left(\frac{W(x)W(a-y)}{W(a)} - I_{(x \geq y)} W(x-y)\right) \Lambda(-\infty, -y)dy. \\ &=: q(a, x, y)dy. \end{aligned} \tag{4.3}$$

Note that the original statement in Bertoin (1997) contains a typo [$\Lambda(-y + dz)$ rather than the correct $\Lambda(dz)$] which is corrected here. Using this proposition, we now derive the asymptotic distribution of (X_{T-}, Δ_T) under the assumption that X_t is of the form (4.1) and that X_t has a positive drift. Under (4.1), the latter assumption is equivalent to

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

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

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

Using Proposition 4.1 we obtain the following result.

Proposition 4.2 *Assume that X_t is compound Poisson as in (4.1) with $\rho < 1$ and that Assumption L holds. Then, as $a \to \infty$, for each x,*

$$P_{a-x}(X_{T-} \in dy;\, \Delta_T \in -dz \mid X_T \le 0) \to \frac{e^{\gamma y} - 1}{1-\rho} \lambda d P(B \le z) dy. \tag{4.4}$$

In particular,

$$P_{a-x}(X_{T-} \in dy \mid X_T \le 0) \to \frac{e^{\gamma y} - 1}{1-\rho} \lambda P(B > y) dy. \tag{4.5}$$

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 (4.2),

$$\begin{aligned} &P_{a-x}(X_{T-} \in dy;\, \Delta_T \in -dz \mid X_T \le 0) \\ &= \frac{W(a-x)W(a-y) - W(a)W(a-x-y)}{W(a) - W(a-x)} \lambda d P(B \le z) dy. \end{aligned}$$

We treat the numerator and denominator on the right hand side of this expression separately. First, we analyze the denominator. Using (2.4), it follows that, as $a \to \infty$,

$$W(a) = \frac{1}{1-\rho} - \frac{C}{1-\rho} e^{-\gamma a}(1 + o(1)). \tag{4.6}$$

This implies, as $a \to \infty$,

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

To obtain the asymptotic behavior of the numerator, we apply (4.6) four times. A simple computation then gives

$$\begin{aligned} &W(a-x)W(a-y) - W(a)W(a-x-y) \\ &= \frac{C e^{-\gamma a}(1 + o(1))}{(1-\rho)^2} \left[1 + e^{\gamma(x+y)} - e^{\gamma x} - e^{\gamma y}\right]. \end{aligned}$$

This implies, as $a \to \infty$,

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

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 $q(a,x,y)dy = P_x(X_{T-} \in dy; X_T \le 0)$.

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

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

Proof A straightforward combination of Proposition 4.1 and (4.6). □

We close this section with some remarks:

- The function $q(x,y)$, appearing as limit in Proposition 4.3 and already defined in Sect. 3, can be interpreted as follows: consider a risk process with initial capital x. Then $q(x,y)dy$ is the probability that ruin eventually occurs, and that the surplus before ruin is in the interval $(y, y+dy)$. The distribution of the surplus prior to ruin has been in investigated in Schmidli (1999).
- 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 (4.1) can be derived from results in Bertoin and Doney (1994), who prove an analogue of (2.4) 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 Assumption L. Recall that

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

By dominated convergence we have $E[N_K] \to E[N] = 1/(1-\rho)$, and from (2.5) we obtain $P(V_{\max} \ge K) \sim C_0 \mathrm{e}^{-\gamma K}$. Thus, to prove Theorem 3.1, it suffices to show that, under Assumption L and $\rho < 1$,

$$E[L_K \mid L_K \geq 1] \to D_0, \tag{5.1}$$

with D_0 defined as in Sect. 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 density $q(a, x, y)$, 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 Sect. 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:

$$p_{K,n}(x, y)dy = P(L_K \geq n; R_K(t_n) \in dy \mid R_K(t_{n-1}) = x; L_K \geq n - 1). \tag{5.2}$$

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

$$p_{K,n}(x, y) = q(K, x, y). \tag{5.3}$$

Set

$$p_{K,n}(y)dy = P(L_K \geq n; R_K(t_n) \in dy \mid L_K \geq 1). \tag{5.4}$$

Then, for $n \geq 2$,

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

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

$$p_{K,1}(y)dy = \int_{u=0}^{K} P_{K-u}(X_{T-} \in dy \mid X_T \leq 0)dP(B \leq u). \tag{5.5}$$

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

$$P(L_K \geq n \mid L_K \geq 1) = \int_0^K p_{K,n}(y)dy. \tag{5.6}$$

We now let $K \to \infty$. Using Proposition 4.3 and (5.3), we obtain

$$p_{K,n}(x, y) \to q(x, y). \tag{5.7}$$

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

$$P_{K-u}(X_{T-} \in dy \mid X_T \leq 0) \to \frac{e^{\gamma y} - 1}{1 - \rho} \lambda P(B > y) dy =: p_1(y) dy. \tag{5.8}$$

It is not difficult to show that for each y, the density of $P_{K-u}(X_{T-} \in dy \mid X_T \leq 0)$ is bounded in K and x, $0 \leq x \leq K$. Thus, using the bounded convergence theorem, we obtain

$$p_{K,1}(y) \to p_1(y). \tag{5.9}$$

From this, we readily obtain by an inductive argument:

$$p_{K,n}(y) \to p_n(y) = \int_{0+}^{\infty} q(x, y) p_{n-1}(x) dx. \tag{5.10}$$

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

$$P(L_K \geq n \mid L_K \geq 1) \to p_n := \int_0^{\infty} p_n(y) dy. \tag{5.11}$$

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

$$E[L_K \mid L_K \geq 1] \to \sum_{n=1}^{\infty} p_n = \sum_{n=1}^{\infty} \int_0^{\infty} p_n(y) dy. \tag{5.12}$$

That this quantity equals D_0 as given by (3.3) can easily be verified by iterating (5.10). 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

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

Since, cf. (2.7),

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

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

$$E[L_K \mid L_K \geq 1] \to 1. \tag{6.1}$$

To prove this, we first assume that inter-arrival times are bounded by a finite constant a, allowing us to use an estimate due to Foss and Zachary (2003). To avoid a notational burden, we do not keep track of a in our notation for the moment.

Since B satisfies Assumption H, there exists a function $h(x) = o(x)$ with $h(x) \to \infty$ as $x \to \infty$ such that $P(B > x) \sim P(B > x - h(x))$, see also Foss and Zachary (2003) for background. Recall that t_1 is the first time a customer gets rejected. Let n_1 be the first customer that gets rejected.

Note that the finite system model and infinite system model coincide up to time t_1. For the infinite system model, let $n(K + a)$ be the first customer experiencing a waiting time larger than $K + a$.

Note that $P(W_{\max} > K + a)$ if and only if $P(n(K + a) \leq N$.

It is shown in Foss and Zachary (2003), that

$$P(W(n(K + a) - 1) > h(K) \mid n(K + a) \leq N) \to 0. \tag{6.2}$$

In words, the customer before customer $n(K + a)$ did not experience a long waiting time. Since inter-arrival times are bounded by a, this implies

$$P(B_{n(K)-1} \leq a + K - h(K) \mid n(K) \leq N) \to 0. \tag{6.3}$$

Note that $P(B_{n(K)-1} > K \mid B_{n(K)-1} > a + K - h(K))$ and consequently,

$$\begin{aligned} P(V_K(t_1-) > h(K) \mid L_K \geq 1) &= P(V(t_1-) > h(K) \mid V_{\max} > K) \\ &\leq P(V(t_1-) > h(K) \mid W_{\max} > K + a) \\ &\leq P(W(n(K + a) - 1) > h(K) \mid W_{\max} > K + a) \\ &\sim P(W(n(K)) > h(K) \mid W_{\max} > K + a) \to 0. \end{aligned}$$

The convergence to 0 follows from a result in Foss and Zachary (2003), while the preceeding equivalence was established in Asmussen (1998)

Now, write

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

We first prove that term I converges to 1 and then show that II converges 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 Bekker and Zwart (2005) 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

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

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 if inter-arrival times are upper bounded by K. For the general case, note that truncating inter-arrival times in a cycle does not decrease the number of losses in a cycle, i.e. $L_K \leq L_K^a$, and apply an argument similar to that in Proposition 2.1. We omit the details.

7 Other rejection mechanisms

In this section we come back to the two other rejection mechanisms mentioned in the Introduction. We present asymptotic expansions for the loss probability of a customer in both cases and use these expansions to compare the various rejection disciplines.

For the $M/G/1$ queue with partial rejection we obtain from (1.2) and (2.5) that, under Assumption L,

$$P_K^p \sim C(E[\mathrm{e}^{\gamma B}] - 1)\mathrm{e}^{-\gamma K} = C\frac{\gamma}{\lambda}\mathrm{e}^{-\gamma K}, \tag{7.1}$$

while under Assumption H,

$$P_K^p \sim E[N]P(B > K). \tag{7.2}$$

The loss probability in the $M/G/1$ queue where customers leave the system impatiently when their waiting time has exceeded K is given by (1.3). Note that the total workload due to *patient* customers satisfies the recursion

$$W_{K,n+1}^i = \begin{cases} (W_{K,n}^i + B_n - A_n)^+ & \text{if } W_{K,n}^i \leq K \\ (W_{K,n}^i - A_n)^+ & \text{if } W_{K,n}^i > K. \end{cases} \tag{7.3}$$

A careful analysis of this recursion may lead to asymptotic expansions for P_K^i as $K \to \infty$ for the $GI/G/1$ queue, but this is not pursued here. Instead, we focus on the $M/G/1$ queue which enables us to apply the exact Formula (1.3). Combining (1.3) with (2.4) we obtain under Assumption L, that

$$P_K^i \sim (1-\rho)C\mathrm{e}^{-\gamma K}. \tag{7.4}$$

Under Assumption H, the residual service time distribution B^r (having density $P(B > x)/E[B]$) is subexponential. This implies, using Pakes (1975) or Veraverbeke (1977),

$$P(W_{M/G/1} > K) \sim \frac{\rho}{1-\rho} P(B^r > x).$$

Combining this result with formula (1.3) then yields

$$P_K^i \sim \rho P(B^r > K). \tag{7.5}$$

From the above expressions one can make the following observations:

- If we compare partial rejection and complete rejection, we observe that

$$\limsup_{K\to\infty} P_K / P_K^p \leq 1$$

for any service-time distribution. This follows by considering the number of rejected customers during a busy cycle. This number is clearly higher under the partial rejection discipline, while the total number of customers entering the system during a busy cycle converges to $1/(1-\rho)$ irrespective of the service discipline. A referee kindly pointed towards the fact that the inequality $P_K/P_K^p \leq 1$ also holds for finite K. For every sample path of the input process, the workload process will be larger under partial rejection than under complete rejection, and therefore every customer which gets (completely) rejected in the latter case, get (partially) rejected in the former case as well. We refrain from adding a detailed inductive proof, which is beyond the scope of this paper.
- A comparison between P_K^i and P_K (or P_K^p) is more difficult to make. If the service times are heavy-tailed it is clear from (7.5) that $P_K^i/P_K \to \infty$; a similar result holds for P_K^i/P_K^p. The reason for this is that the first overshoot of level K in a busy cycle grows without bound as $K \to \infty$, cf. a result of Asmussen and Klüppelberg (1997). Thus, for the model with impatience it follows that, when the first loss in a cycle occurs, it takes a long time to return to level K (cf. the recursion 7.3). This leads to many abandonments in a cycle implying that P_K^i is intrinsically larger than $P(V_{\max} > K)$.
- If we compare the asymptotics (7.4) and (7.1) under Assumption L, the above-mentioned heuristics do not apply—in fact it follows from the results in Sect. 4 that the first overshoot of level K in a busy cycle does converge to a proper distribution. Thus, it may happen that P_K^p can be asymptotically larger than P_K^i. In particular, it follows from (7.1) and (7.4) that

$$\lim_{K\to\infty} P_K^p/P_K^i = \frac{\gamma}{\lambda(1-\rho)}.$$

For the $M/M/1$ queue we have $\gamma = \mu(1-\rho)$, so that the above ratio becomes $1/\rho > 1$.

- A comparison of P_K and P_K^i under Assumption L does not lead to simple results. For the $M/M/1$ queue for example, we can show that, using (1.4), $P_K/P_K^i \to \mathrm{e}^{-\rho}/\rho$, which is smaller (resp. larger) than 1 if ρ is large (small). For the $M/D/1$ queue we have the identity $P_K = P_{K-D}^i$ which, since P_K^i is decreasing in K, implies $P_K \geq P_K^i$.

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

References

Asmussen S, Perry D (1995) Rejection rules in the $M/G/1$ queue. Queueing Syst 19:105–130
Asmussen S, Sigman K (1996) Monotone stochastic recursions and their duals. Probab Eng Inf Sci 10:1–20
Asmussen S, Klüppelberg C (1997) Stationary $M/G/1$ excursions in the presence of heavy tails. J Appl Probab 33:208–212
Asmussen S (1998) Subexponential asymptotics for stochastic processes: extremal behaviour, stationary distributions and first passage times. Ann Appl Probab 8:354–374
Asmussen S (2003) Appl Probab Queues, 2nd edn. Springer, New York
Asmussen S, Pihlsgaard M (2007) Loss rates for Lévy processes with two reflecting barriers. Math Oper Res 32:308–321
Bekker R, Zwart AP (2005) On an equivalence between loss rates and cycle maxima in queues and dams. Probab Eng Inf Sci 19:241–255
Bertoin J, Doney R (1994) Cramér's estimate for Lévy processes. Stat Probab Lett 21:363–365
Bertoin J (1997) Exponential Decay and Ergodicity of completely asymmetric Lévy processes on a finite interval. Ann Appl Probab 7:156–169
Boots NK, Tijms HC (1999) A multiserver queueing system with impatient customers. Manag Sci 45:444–448
Cohen JW (1969) Single-server queues with restricted accessibility. J Eng Math 3:265–284
Cohen JW (1976) Regen processes queueing theory. Springer, Berlin
Cohen JW (1982) The single server queue. North Holland, Amsterdam
Debicki K, Mandjes M (2012) Lévy-driven queues. Surv Oper Res Manag Sci 17:15–37
Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling extremal events. Springer, Berlin
Foss S, Zachary S (2003) The maximum on a random time interval of a random walk with long-tailed increments and negative drift. Ann Appl Probab 13:37–53
Gavish B, Schweitzer P (1977) The Markovian queue with bounded waiting time. Manag Sci 23:1349–1357
Iglehart DG (1972) Extreme values in the $GI/G/1$ queue. Ann Math Stat 43:627–635
de Kok AG, Tijms HC (1985) A two-moment approximation for a buffer design problem requiring a small rejection probability. Perform Eval 5:77–84
Kyprianou AE, Palmowski Z (2004) A martingale review of some fluctuation theory for spectrally negative Lévy processes. Sémin Probab XXXVIII:16–29
van Ommeren JCW (1987) Exponential bounds for excess probabilities in systems with a finite capacity. Stoch Process Appl 24:143–149
Norvang Andersen L (2011) Subexponential loss rate asymptotics for Lévy processes. Math Methods Oper Res 73:91–108
Pakes AG (1975) On the tails of waiting-time distributions. J Appl Probab 12:555–564
Schmidli H (1999) On the distribution of the surplus prior and at ruin. ASTIN Bull 29:227–244
Suprun VN (1976) Problem of destruction and resolvent of terminating processes with independent increments. Ukr Math J 28:39–45
Takács L (1967) Combinatorial methods in the theory of stochastic processes. Wiley, New York
Tijms HC (2003) A first course in stochastic models. Wiley, New York

Veraverbeke N (1977) Asymptotic behaviour of Wiener-Hopf factors of a random walk. Stoch Process Appl 5:27–37
Zwart AP (2000) A fluid queue with a finite buffer and subexponential input. Adv Appl Probab 32:221–243