Adv. Appl. Prob. **32,** 1150–1167 (2000)
Printed in Northern Ireland

OVERFLOW BEHAVIOR IN QUEUES WITH MANY LONG-TAILED INPUTS

MICHEL MANDJES,* *Bell Laboratories, Lucent Technologies*
SEM BORST,** *CWI*

Abstract

We consider a fluid queue fed by the superposition of n homogeneous on–off sources with generally distributed on and off periods. The buffer space B and link rate C are scaled by n, so that we get nb and nc, respectively. Then we let n grow large. In this regime, the overflow probability decays exponentially in the number of sources n. We specifically examine the scenario where b is also large. We obtain explicit asymptotics for the case where the on periods have a subexponential distribution, e.g., Pareto, Lognormal, or Weibull.

The results show a sharp dichotomy in the qualitative behavior, depending on the shape of the function $v(t) := -\log \mathrm{P}(A^\star > t)$ for large t, $A^\star$ representing the *residual* on period. If $v(\cdot)$ is regularly varying of index 0 (e.g., Pareto, Lognormal), then, during the path to overflow, the input rate will only slightly exceed the link rate. Consequently, the buffer will fill 'slowly', and the typical time to overflow will be 'more than linear' in the buffer size. In contrast, if $v(\cdot)$ is regularly varying of index strictly between 0 and 1 (e.g., Weibull), then the input rate will significantly exceed the link rate, and the time to overflow is roughly proportional to the buffer size.

In both cases there is a substantial fraction of the sources that remain in the on state during the entire path to overflow, while the others contribute at their mean rates. These observations lead to approximations for the overflow probability. The approximations may be extended to the case of heterogeneous sources. The results provide further insight into the so-called reduced-load approximation.

Keywords: Buffer overflow; large-deviations asymptotics; long-tailed on periods; long-range dependence; on–off sources; queueing theory; reduced-load approximation; regular variation; subexponentiality

AMS 2000 Subject Classification: Primary 60K25
Secondary 68M20; 90B18; 90B22

1. Introduction

Measurements have indicated that network traffic exhibits burstiness on a wide range of time scales [27]. This conclusion was drawn after thorough examination of traffic streams in a variety of packet-based networks, see for instance [6], [38]. The discovery of the presence of *long-range dependence* had a major impact on traffic modeling. Where one used to rely on short-range dependent models, recent work has witnessed an increasing interest in traffic models which exhibit burstiness on a wider range of time scales.

Received 15 November 1999; revision received 17 May 2000.
* Postal address: Bell Laboratories, 600 Mountain Avenue, P.O. Box 636, Murray Hill, NJ 07974, USA.
** Postal address: CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands. Email address: sem@cwi.nl

The crucial characteristic of long-range dependent traffic is that it does not obey a Markovian correlation structure, as such a structure is inherently short-range dependent. Several models have been proposed to capture the essential features. As described in the survey by Boxma and Dumas [11], three major approaches may be distinguished. (i) Erramilli *et al.* [19] advocate the application of *chaotic maps*. (ii) Norros [34], [35] proposes the use of *fractional Brownian motion*. (iii) Willinger *et al.* [46] use the superposition of on–off sources with *long-tailed* activity periods. In the present paper, we follow the third approach.

An on–off source alternates between activity periods (commonly referred to as bursts) and silence periods. On–off models have appeared to be extremely versatile. Specific choices for the distributions of the on and off periods enable us to capture the relevant characteristics of network traffic. In the basic model, a superposition of these sources feeds into a buffer which is emptied at constant rate C. One then often focuses on the probability of the buffer content exceeding some level B.

Kosten [25], [26], Anick *et al.* [3], Elwalid and Mitra [18] and Stern and Elwalid [44] examine the case where the on and off periods are mixtures of exponential distributions. They explicitly find both the steady-state buffer content distribution and large-buffer asymptotics. In these models, the overflow probability decays essentially exponentially in the buffer level B.

In the present paper, we depart from the assumption that the on and off periods have exponentially bounded tails. Instead, we assume 'long-tailed' distributions, such as Pareto, Lognormal, and Weibull. Based on results in the literature discussed below, we expect that long-tailed activity periods should cause the overflow probability to decay slower than exponentially in the buffer level B.

1.1. Literature

The early literature on long-tailed queues goes back to the 1970s. Important contributions were made by Cohen [12] and Pakes [36]. They consider $GI/G/1$ queues in which the residual service time has a subexponential distribution. The class of subexponential distributions is an important subclass of the class of long-tailed distributions, see Section 2. Remarkably, the waiting-time distribution has a similar shape to the residual service time.

During the 1990s the focus shifted to on–off sources. Boxma [9] and Jelenković and Lazar [23] consider a queue fed by a single source. Applying the result for the $GI/G/1$ queue, they obtain the large-buffer asymptotics. It turns out the tail behavior is mainly determined by the distribution of the residual on period — provided this random variable has a subexponential distribution. This result will play a crucial role in our analysis.

The analysis in the case of *multiple* on–off sources is fundamentally more complicated. Dumas and Simonian [17] derive asymptotic upper and lower bounds for the buffer content distribution. Agrawal *et al.* [1] consider the special case of two on–off sources. Under certain conditions, the buffer content is proved to be asymptotically equivalent to that in a reduced system. The reduced system consists of a queue fed by only the 'heavier-tailed' one of the two sources, emptied at the link rate C subtracted by the mean rate of the other source. This *reduced-load equivalence* had also been found by Boxma [10] and Jelenković and Lazar [23] in the case of one regularly varying source and several exponential sources. The result requires the assumption that the peak rate of the heavy-tailed source, increased by the mean rates of the other sources, exceeds the link rate C. In a recent paper, Zwart *et al.* [48] extend the reduced-load equivalence, and obtain the exact large-buffer asymptotics for the general case of multiple heterogeneous on–off sources with regularly varying on periods.

A related class of models is that of $M/G/\infty$ input. Sessions arrive as a Poisson process, and remain in the system for a generally distributed time, during which they generate traffic at a fixed rate. Resnick and Samorodnitsky [39] derive large-buffer asymptotics for regularly varying holding times. Parulekar and Makowski [37] and Duffield [15] obtain large-buffer asymptotics for subexponential holding times. In [15] the Poisson arrival rate, link rate, and buffer size are scaled as $\Lambda \equiv n\lambda$, $C \equiv nc$ and $B \equiv nb$, respectively, with n growing large. This regime allows explicit asymptotic analysis, as results from large-deviations theory become applicable. Related results may be found in [30] and [31]. Likhanov and Mazumdar [29] generalize the large-buffer asymptotics for regularly varying holding times to the case of heterogeneous session characteristics.

A relevant performance measure is also the expected time until overflow of a given large buffer level. Heath *et al.* [21] show that this quantity is strongly affected by the values of the system load and the rate of individual sessions relative to the link rate. Resnick and Samorodnitsky [40] give the exact asymptotics of this quantity for the case where a single long active session is sufficient to cause a positive drift in the buffer content.

1.2. Contributions

Following Duffield [15], we scale buffer and link rate, in our case with the number of sources n. We focus on the case where the exponentiality assumptions on bursts and silence periods are removed. Applying large-deviations techniques, we show that the overflow probability decays exponentially in the scaling parameter n, with decay rate $I(b)$ as a function of the scaled buffer size b. Within this regime, we are interested in large-buffer asymptotics. In other words, we examine $I(b)$ for large b. In this setting the present paper makes the following two contributions.

- We find a function $v(\cdot)$ such that $I(b)/v(b)$ tends to a positive constant, for large b. This is done by characterizing the moment generating function of the traffic generated by a single source in an interval of length t, under a specific scaling. The scaling was first proposed by Parulekar and Makowski [37] for the $M/G/\infty$ case, but also applies in our model. Then we exploit this characterization to establish the asymptotics of $I(b)$ for large b. As mentioned above, all large-buffer asymptotics obtained previously require specific model assumptions; ours do not. The trade-off is that the results are asymptotic in the number of sources as well as the buffer size.
  In particular, we show that if the residual on period is subexponential, then so is the buffer content distribution (i.e., the growth of $I(b)$ is slower than linear). The practical implication is that buffer dimensioning based on Markovian models (for which $I(b)$ is essentially a straight line) would be overly optimistic in the case of large buffers. This result may be seen as complementary to that in [32]. There it is shown that, in the case of small buffers, long-range dependence does not have a substantial effect on the loss probability. We refer the reader to [20], [22] and [42] for related results indicating the significant influence of the buffer size on the impact of long-range dependence.
- We contribute to the understanding of *the way overflow occurs*. Clearly, to fill a large buffer, there is a trade-off between the *intensity* of the deviant behavior (to what extent does the input rate exceed the link rate?) and the *duration* of the deviant behavior. For on–off sources with exponentially bounded on periods, it is known that sources alternate between on and off during the path to overflow, but with longer on periods and shorter off periods; all sources behave in essentially the same way [2], [33], [45]. In contrast, if the on periods are subexponential, then sources contribute either at their mean rate or

their peak rate. Put differently, some sources remain in the on state during the entire path to overflow, while the others alternate between on and off (thus effectively contributing at their mean rates). We can explicitly calculate the number of sources that transmit at peak rate all the time.

This understanding is exploited to derive a number of approximations for the overflow probability and also for the case of heterogeneous sources. For regularly varying on periods, this yields a reduced-load equivalence, which is in agreement with the bounds of Dumas and Simonian [17].

1.3. Organization

The remainder of the paper is organized as follows. Section 2 describes the model, introduces notation, and gives some basic definitions and assumptions. Section 3 presents the analysis. We establish the structure of the cumulant function of the traffic generated by a single source, under a critical scaling. Section 4 concentrates on the intuition behind the results and provides qualitative insights. Section 5 concludes.

2. Model and preliminaries

In the first subsection, we introduce the model and the required assumptions. The second subsection briefly reviews the class of subexponential distributions and states the asymptotic result for a queue fed by a single source with subexponential on periods.

2.1. Model description

We consider traffic from n on–off sources arriving at a buffered resource. The resource is modeled as a queue with constant depletion rate C. The traffic rate of each source alternates between a peak rate r and 0. The activity periods form an i.i.d. sequence of random variables, each of them distributed as random variable A. We assume that A has unbounded support. The silence periods are also an i.i.d. sequence, distributed as random variable S. Both sequences are mutually independent. We also define $A(t)$ to be the traffic generated by a single source in steady state in the time interval $[0, t]$. Later in our analysis we need the following assumption.

Assumption 2.1. *The random variables A and S are such that $\mathrm{E}A^{1+\zeta} < \infty$ (for some positive ζ) and $\mathrm{E}S < \infty$. The distribution of $A + S$ is non-lattice.*

The above assumption has two major implications — for details we refer the reader to Section 2.1 of [17]. In the first place, the fact that both $\mathrm{E}A$ and $\mathrm{E}S$ are finite ensures that the long-run fraction of time that the source spends in the on state is

$$p := \frac{\mathrm{E}A}{\mathrm{E}A + \mathrm{E}S},$$

and the fraction of time spent in the off state is the complement, $1 - p$. Also, the residual activity period $A^\star$ is well defined: conditioned on the process being in the on state, $A^\star$ has distribution

$$F_{A^\star}(x) := \mathrm{P}(A^\star > x) = \frac{1}{\mathrm{E}A} \int_x^\infty \mathrm{P}(A > y)\, \mathrm{d}y.$$

We are interested in the probability of the buffer content exceeding some level B, denoted by $p(B, C)$. We rescale the resources by the number of sources: $C \equiv nc$ and $B \equiv nb$. This scaling was first introduced by Weiss [45] and has proved to be very powerful (see, for instance,

3. Analysis

We focus on the situation with a large number of sources feeding into a large buffer. As mentioned in the introduction, we investigate the asymptotics of the exponential decay rate $I(b)$ for large values of the buffer level b.

In the first subsection, we review the relevant large-deviations results, which enable us to calculate $I(b)$ for general b. This expression remains somewhat implicit: it turns out to be the solution to a variational problem. In particular, we concentrate on the conditions under which this result applies. We also stress the role of the so-called 'scaling function'.

In the second subsection, we study the logarithm of the moment generating function of the random variable $A(t)$, also called the *cumulant function*. The cumulant function is needed in the variational problem mentioned above. We show that, under a specific scaling, the cumulant function is piecewise linear for on–off sources with subexponential on periods. The choice of the particular scaling is due to Duffield [15] and Parulekar and Makowski [37].

The third subsection contains the main result: the asymptotics (for large b) of the decay rate. We combine the variational problem of Subsection 3.1 and the cumulant function of Subsection 3.2. Here, we follow the approach of Duffield [15] for the $M/G/\infty$ case.

3.1. Decay rate for general buffer level

In this subsection, we focus on the evaluation of the decay rate for general buffer level b. The theorem which we will use in Subsection 3.3 is a variant of the key theorem of [28], which is stated in Theorem 3.2 below. In [28], this result is phrased in the setting of slotted time; in [32] it is extended to continuous time. The latter version is formulated in Theorem 3.2 below, and requires the following assumption.

Assumption 3.1. *Define*

$$J_t(x) := \sup_\theta(\theta x - \log \mathrm{E}\, \mathrm{e}^{\theta A(t)})$$

and assume that:

(i) *for any* $b \geq 0$, $\liminf_{t\to\infty} J_t(b + ct)\,(\log t)^{-1} > 0$;

(ii) $J(b) := \inf_{t>0} J_t(b + ct)$ *is a continuous function of* b.

Theorem 3.2. (Loss curve for general b.) *Under Assumption 3.1,*

$$I(b) := -\lim_{n\to\infty} \frac{1}{n} \log p_n(b, c) = J(b) = \inf_{t>0} \sup_\theta(\theta(b + ct) - \log \mathrm{E}\, \mathrm{e}^{\theta A(t)}). \tag{1}$$

For the proof of Theorem 3.2, we refer the reader to [32]. Assumption 3.1(ii) is of a technical nature, and will be satisfied in all cases of practical interest. Assumption 3.1(i) is due to Likhanov and Mazumdar [28]. In earlier versions of Theorem 3.2, e.g., the one in [8], the conditions imposed on the input process were usually more restrictive. In [8] it is required that there is a θ such that for t large enough $\log \mathrm{E}\, \mathrm{e}^{\theta A(t)} < c\theta t$.

It is not hard to show that this condition is not fulfilled for on–off sources with heavy-tailed on periods (see [32]). The (weaker) requirement Assumption 3.1(i) is satisfied under the non-restrictive condition that $\mathrm{E}A^{1+\zeta}$ is finite for some positive ζ, as we postulated in Assumption 2.1. We prove this in the next proposition.

[8], [13] and [43]). We assume the system is stable and non-trivial:

$$\rho := pr < c < r.$$

In the scaled model we define $p_n(b, c)$ to be the steady-state probability that the buffer content exceeds level nb. In particular, we will analyze the exponential *decay rate* (as a function of b, for fixed c):

$$I(b) := -\lim_{n\to\infty} \frac{1}{n} \log p_n(b, c),$$

given that the limit exists.

2.2. Subexponentiality

In this subsection, we give the definition of subexponential distributions and state the asymptotic result for a queue fed by a single source with subexponential on periods. More details may be found in the appendices of [11]. Throughout, we denote by $F_X(\cdot)$ the distribution function of the random variable X, with $F_{A^*}(\cdot)$ specifically indicating the distribution function of the residual activity period.

Definition 2.2. (*Subexponential distribution.*) Suppose that

$$\frac{\mathrm{P}(X + X' > t)}{\mathrm{P}(X > t)} \to 2, \qquad t \to \infty,$$

where X and X' are i.i.d. random variables. Then we say that X has a *subexponential distribution*, or $F_X(\cdot) \in \mathcal{S}$.

Besides $\mathcal{S}$, we introduce a second class of distribution functions. In this class, a crucial role is played by the function $v_X(t) := -\log \mathrm{P}(X > t)$. In Section 3 we will show that the shape of this function determines the large-buffer asymptotics, with $X = A^*$.

Definition 2.3. (*Subexponentially varying distribution.*) Suppose that the function $v_X(\cdot)$ is regularly varying of index h (at infinity), that is,

$$\frac{v_X(yt)}{v_X(t)} \to y^h, \qquad t \to \infty,$$

for all $y > 0$. If $v_X(\cdot)$ is regularly varying of index $h \in [0, 1)$, then we say that X has a *subexponentially varying distribution*, or $F_X(\cdot) \in \mathcal{V}$.

In the last definition we used the concept of regular variation (see, for instance, Section 1.4 of [7]). The exact relationship between the classes $\mathcal{S}$ and $\mathcal{V}$ is not clear. However, the most important long-tailed distributions (such as Pareto, Lognormal, or Weibull) are in both of them.

The following theorem is an extension of the results for the $GI/G/1$ queue in [12] and [36], and may be found in [9] and [23]. This result for a queue fed by a single source will be one of the main building blocks in the analysis of the queue fed by n sources.

Theorem 2.4. (Single source.) *Denote by Q the steady-state buffer content in a queue with service rate c fed by a single source with $\rho < c < r$. If $F_{A^*}(\cdot) \in \mathcal{S}$, then*

$$\mathrm{P}(Q > b) \sim (1 - p)\frac{\rho}{c - \rho}\mathrm{P}\left(A^* > \frac{b}{r - c}\right),$$

where '$\sim$' means that the ratio of both sides tends to 1 as $b \to \infty$.

Proposition 3.3. *Consider an on–off source with on periods A.*

(i) *Assumption 3.1(i) is satisfied if* $\mathrm{E}A^{1+\zeta} < \infty$ *for some positive* ζ.

(ii) *If* $F_{A^*}(\cdot) \in \mathcal{S}$, *then for any positive* $\varepsilon < r - \rho$, *there is a positive constant* $K_{\delta\varepsilon}$ *such that for t large enough*

$$\mathrm{P}\left(\frac{A(t)}{t} > \rho + \varepsilon\right) \leq K_{\delta\varepsilon}\mathrm{P}(A^* > \delta^* t) \quad \text{where } \delta^* := \frac{(1-\delta)\varepsilon}{r - (\rho + \delta\varepsilon)},\ \delta \in (0,1).$$

Proof. We start with the second statement.

(ii) We may write

$$\begin{aligned}\mathrm{P}\left(\frac{A(t)}{t} > \rho + \varepsilon\right) &= \mathrm{P}(A(t) - (\rho + \delta\varepsilon)t > (1-\delta)\varepsilon t)\\ &\leq \mathrm{P}(\exists s : A(s) - (\rho + \delta\varepsilon)s > (1-\delta)\varepsilon t)\\ &= \mathrm{P}(Q > (1-\delta)\varepsilon t),\end{aligned}$$

where Q is defined to be the steady-state buffer content when the source feeds into a queue with service rate $\rho + \delta\varepsilon$. Using Theorem 2.4, we see that there is a $K_{\delta\varepsilon}$ such that for t large enough the statement holds.

(i) Proposition 3.1 of [28] states that Assumption 3.1(i) is satisfied if for all $\varepsilon \in (0, r-\rho)$ there are positive K and α such that for t large enough

$$\mathrm{P}\left(\frac{A(t)}{t} > \rho + \varepsilon\right) \leq Kt^{-\alpha}.$$

As above,

$$\mathrm{P}\left(\frac{A(t)}{t} > \rho + \varepsilon\right) \leq \mathrm{P}(Q > \tfrac{1}{2}\epsilon t),$$

where Q denotes the steady-state buffer content when the source feeds into a queue with service rate $\rho + \frac{1}{2}\varepsilon$. If $F_{A^*}(\cdot) \in \mathcal{S}$ (which we will assume in most of the sequel anyway), then the desired statement follows as above, since $\mathrm{E}A^{1+\zeta} < \infty$ implies that $\mathrm{P}(A^* > \delta^* t) \leq t^{-\zeta}$ for t large enough.

To see that the statement also holds when $F_{A^*}(\cdot) \notin \mathcal{S}$ (in which case Theorem 2.4 cannot be used), we may invoke a result of [24] to relate the buffer content in a fluid queue to the waiting time W in a corresponding $GI/G/1$ queue with service times proportional to A. Theorem 2.1 of Chapter 8 of [4] states that $\mathrm{E}W^{\zeta} < \infty$ if $\mathrm{E}A^{1+\zeta} < \infty$. Using Markov's inequality, the desired statement then follows directly with $\alpha = \zeta$.

Corollary 3.4. *An alternative variational problem to compute the decay rate is given by*

$$I(b) := -\lim_{n\to\infty} \frac{1}{n} \log p_n(b,c) = \inf_{t>0} w(t)\bar{J}_t\left(\frac{b}{t} + c\right),$$

where

$$\bar{J}_t(x) := \sup_{\theta}\left(\theta x - \frac{\log \mathrm{E}\,\mathrm{e}^{\theta A(t)w(t)/t}}{w(t)}\right), \tag{2}$$

for an increasing positive function $w(\cdot)$. *Compared to the variational problem of (1), the optimum is attained at the same* t^* *and a value of* θ^* *that is* $t^*/w(t^*)$ *times as large.*

The function $w(\cdot)$ in Corollary 3.4 is usually called a *scaling function*. It was introduced in [14], [16] to enable large-deviations analysis in situations where there is no exponential decay. For the case of $M/G/\infty$ input, the use of the scaling $w(t) = -\log \mathrm{P}(D^\star > t)$, with $D^\star$ representing the residual session length was proposed in [37]. In the sequel, we use the scaling

$$w(t) = v(t) := v_{A^\star}(t) = -\log \mathrm{P}(A^\star > t).$$

3.2. The cumulant function

As observed in Subsection 3.1, the moment generating function of $A(t)$ plays a crucial role in determining the decay rate $I(b)$. In view of Corollary 3.4, we are interested in the asymptotic behavior of

$$\frac{\log \mathrm{E}\, \mathrm{e}^{\theta A(t)v(t)/t}}{v(t)}. \tag{3}$$

In this subsection, we prove that, for t large, this cumulant function is piecewise linear in θ. The exact statement is given in Theorem 3.6, but we provide an intuitive explanation for the result first.

The intuition for queues with heavy-tailed activity periods is that — during the path to overflow — with overwhelming probability, a source sends either at mean rate, or sends *essentially the entire time interval* at peak rate. This behavior is reflected by the following asymptotics: in Theorem 3.6 we will prove that

$$\begin{aligned}\mathrm{E} \exp(\theta A(t)v(t)/t) &\approx (1 - \mathrm{P}(A^\star > t))\, \mathrm{e}^{\theta \rho v(t)} + \mathrm{P}(A^\star > t)\, \mathrm{e}^{\theta r v(t)} \\ &\approx \exp[v(t) \max\{\theta\rho, \theta r - 1\}],\end{aligned}$$

as $t \to \infty$. In order to prove (the formal version of) this statement, we first establish an auxiliary lemma.

Lemma 3.5. *For all θ, with $h \in [0, 1)$,*

$$\max_{x \in [0, r-\rho]} f_x(\theta) = \max\{\theta\rho, \theta r - 1\}, \quad \textit{where } f_x(\theta) := \theta(\rho + x) - \left(\frac{x}{r-\rho}\right)^h.$$

Here, x^h with $h = 0$ is defined to be 1 for $x > 0$, and to be 0 for $x = 0$.

Proof. The proof follows directly from the convexity of $f_x(\theta)$ in x. However, we give an alternative proof to provide additional insight. First note that, if $x = 0$, we get curve $\theta\rho$; for $x = r - \rho$ we get $\theta r - 1$.

Now take an x in the interior of the interval: $x \in (0, r - \rho)$. The lines $\theta\rho$ and $\theta r - 1$ intersect at $\theta_0 := (r - \rho)^{-1}$. Suppose that we can prove that

$$f_x(\theta_0) = \theta(\rho + x) - \left(\frac{x}{r-\rho}\right)^h \Big|_{\theta=\theta_0} < \max\{\theta_0\rho, \theta_0 r - 1\} = \frac{\rho}{r-\rho}, \tag{4}$$

then we are done. This is because the slope of $f_x(\theta)$ (as a function of θ) is in the interval (ρ, r). So, if (4) holds, then $f_x(\theta_0)$ is smaller than or equal to $\theta\rho$ for $\theta \le \theta_0$, and smaller than or equal to $\theta r - 1$ for $\theta \ge \theta_0$.

Statement (4) follows directly from the standard algebraic inequality

$$\theta_0(\rho + x) - \left(\frac{x}{r-\rho}\right)^h < \theta_0(\rho + x) - \left(\frac{x}{r-\rho}\right) = \frac{\rho}{r-\rho},$$

recalling that $h \in [0, 1)$ and $x \in (0, r - \rho)$.

Theorem 3.6. (Cumulant function.) *For on–off sources with* $F_{A^*}(\cdot) \in \mathcal{S} \cap \mathcal{V}$ *and* $\theta \geq 0$,

$$\lim_{t\to\infty} (\log \mathrm{E} \exp(\theta A(t)v(t)/t))/v(t) = \max\{\theta\rho, \theta r - 1\}, \tag{5}$$

with $v(\cdot)$ *regularly varying of index* $h \in [0, 1)$.

Proof. The proof consists of (i) a lower bound and (ii) a matching upper bound.

(i) Lower bound. The limit in the left-hand side of (5) is larger than $\theta\rho$ due to Jensen's inequality. This bound holds for all $\theta \geq 0$.

A second lower bound can be found by considering the event that the source remains in the on state during the entire interval $[0, t]$. For all $\theta \geq 0$, $t > 0$:

$$\mathrm{E} \exp(\theta A(t)v(t)/t) \geq p\mathrm{P}(A^\star > t)\, \mathrm{e}^{\theta r v(t)} = p\, \mathrm{e}^{(\theta r - 1)v(t)}, \tag{6}$$

with p denoting the probability of the on state.

Combining both lower bounds yields the desired result:

$$\liminf_{t\to\infty} (\log \mathrm{E} \exp(\theta A(t)v(t)/t))/v(t) \geq \max\{\theta\rho, \theta r - 1\}.$$

(ii) Upper bound. First take $\theta \geq 0$. Choose a $k \in \mathbb{N}$ and $\varepsilon := \varepsilon_k = (r - \rho)/k$. Then

$$\mathrm{E}\, \mathrm{e}^{\theta A(t)v(t)/t} \leq \mathrm{e}^{\theta(\rho+\varepsilon)v(t)} + \sum_{i=1}^{k-1} \mathrm{e}^{\theta(\rho+(i+1)\varepsilon)v(t)} \mathrm{P}\left(\frac{A(t)}{t} \in [\rho + i\varepsilon, \rho + (i+1)\varepsilon]\right).$$

Now use Proposition 3.3(ii). For any $\delta \in (0, 1)$

$$\mathrm{P}\left(\frac{A(t)}{t} \in [\rho + i\varepsilon, \rho + (i+1)\varepsilon]\right) \leq K_{i\delta\varepsilon} \exp\left[-v\left(\frac{(1-\delta)i\varepsilon}{r - (\rho + \delta i\varepsilon)} t\right)\right].$$

Using the fact that $v(\cdot)$ is regularly varying of index h, we obtain

$$\limsup_{t\to\infty} \frac{\log \mathrm{E} \exp(\theta A(t)v(t)/t)}{v(t)} \leq \max_{i=0,\ldots,k-1} \left\{\theta(\rho + (i+1)\varepsilon) - \left(\frac{(1-\delta)i\varepsilon}{r - (\rho + \delta i\varepsilon)}\right)^h\right\}.$$

Optimizing over a continuous, rather than a discrete, domain gives the upper bound

$$\max_{x\in[0,r-\rho]} \left\{\theta(\rho + x + \varepsilon) - \left(\frac{(1-\delta)x}{r - (\rho + \delta x)}\right)^h\right\}.$$

Now let $\delta \downarrow 0$, $k \to \infty$ (and hence $\varepsilon_k \downarrow 0$) and use Lemma 3.5. The upper bound then follows.

3.3. Large-buffer asymptotics of the decay rate

In this subsection, we combine the large-deviations results for general buffer level b, as were obtained in Subsection 3.1, and the specific structure of the cumulant function, as derived in Subsection 3.2. The proof is similar to that of Duffield [15] for the $M/G/\infty$ case. We first prove the analogue of Duffield's Lemma 6.

Lemma 3.7. *The following statements hold for* $\bar{J}_t(x)$, $t \to \infty$:

(i) *For all* $x \in (\rho, r)$, $\lim_{t\to\infty} \bar{J}_t(x) = (x - \rho)/(r - \rho)$.

(ii) *The convergence in (i) is uniform on compact subsets of* (ρ, r).

Proof. (i) Notice that $\bar{J}_t(x)$ is the Legendre–Fenchel transform of the cumulant function (3). The following result is established in the proof of Theorem 2 of [14]. Let f_t be a sequence of convex functions that converge pointwise to f on the interior of the effective domain of f. Then the Legendre–Fenchel transforms $f_t^*(x) := \sup_\theta(\theta x - f_t(\theta))$ also converge to $f^*(x) := \sup_\theta(\theta x - f(\theta))$ on the interior of the effective domain of f^*. Therefore, for $x \in (\rho, r)$,

$$\lim_{t\to\infty} \bar{J}_t(x) = \sup_\theta \left(\theta x - \lim_{t\to\infty} \frac{\log \mathrm{E}\, \mathrm{e}^{\theta A(t) v(t)/t}}{v(t)} \right).$$

From the facts that $\mathrm{E}A(t) = \rho t$ and $x \in (\rho, r)$, it easily follows that the above supremum needs to be taken over positive θ only. Using Theorem 3.6 and observing that

$$\sup_{\theta>0}(\theta x - \max\{\theta\rho, \theta r - 1\}) = \frac{x-\rho}{r-\rho}$$

if $x \in (\rho, r)$, we are done.

(ii) As may be found in Theorem 10.8 of [41], the following property holds: if finite convex functions f_t converge pointwise to a finite convex function f on a certain domain, then the convergence is uniform on compact subsets of the domain.

So it remains to prove that $\bar{J}_t(x)$ is finite for $x \in (\rho, r)$. Take an x in this interval; as above, the supremum in (2) needs to be taken over positive θ only. We arrive at

$$\begin{aligned} \sup_{\theta>0} \left(\theta x - \frac{\log \mathrm{E}\, \mathrm{e}^{\theta A(t) v(t)/t}}{v(t)} \right) &\le \sup_{\theta>0} \left(\theta x - \frac{\log(p\mathrm{P}(A^* > t)\, \mathrm{e}^{\theta r v(t)})}{v(t)} \right) \\ &= \sup_{\theta>0} \left(\theta x - \frac{\log p}{v(t)} - \theta r + 1 \right) < \infty, \end{aligned}$$

for $x \in (\rho, r)$, where the first inequality follows from (6).

We are now in a position to prove the main theorem. It states that for on–off sources with $F_{A^*}(\cdot) \in \mathcal{V} \cap \mathcal{S}$, the loss curve $I(b)$ is, up to a multiplicative constant, asymptotically equal to $v(b)$. In other words, we come to the remarkable conclusion that the residual on period completely determines the large-buffer asymptotics; the off-period distribution contributes only by its mean, via ρ.

Theorem 3.8. (Large-buffer asymptotics.) *The following large-buffer asymptotics of the decay rate hold for on–off sources with $F_{A^*}(\cdot) \in \mathcal{S} \cap \mathcal{V}$:*

$$\lim_{b\to\infty} \frac{I(b)}{v(b)} = \begin{cases} \dfrac{c-\rho}{r-\rho} & \text{if } h = 0, \\ \dfrac{c-\rho}{r-\rho} \dfrac{1}{1-h} \left(\dfrac{h}{1-h}(c-\rho) \right)^{-h} & \text{if } h \in (0,1) \text{ and } \dfrac{c-\rho}{r-\rho} \dfrac{1}{1-h} \le 1, \\ \left(\dfrac{1}{r-c} \right)^h & \text{if } h \in (0,1) \text{ and } \dfrac{c-\rho}{r-\rho} \dfrac{1}{1-h} > 1. \end{cases}$$

Proof. The proof consists of (i) an upper bound and (ii) a matching lower bound.

(i) Upper bound. The proof of the upper bound is parallel to that in [15] for the $M/G/\infty$ case. By Corollary 3.4,

$$\limsup_{b\to\infty}\frac{I(b)}{v(b)}=\limsup_{b\to\infty}\frac{\inf_{t>0}v(t)\bar{J}_t(b/t+c)}{v(b)}.$$

Note that we in fact minimize over $t\geq b\,(r-c)^{-1}$, as the rate function is infinite for smaller t. A formal proof of this statement is not difficult; the intuition is that t represents the time to overflow starting in an empty system; this must obviously be later than $b\,(r-c)^{-1}$.

Substituting $s:=b/t$, we have for all $s\in(0,r-c)$

$$\limsup_{b\to\infty}\frac{I(b)}{v(b)}\leq\limsup_{b\to\infty}\frac{v(b/s)\bar{J}_{b/s}(s+c)}{v(b)}=s^{-h}\frac{s+c-\rho}{r-\rho},$$

where the last equality follows from Lemma 3.7(i). As this holds for all $s\in(0,r-c)$,

$$\limsup_{b\to\infty}\frac{I(b)}{v(b)}\leq\inf_{s\in(0,r-c)}s^{-h}\frac{s+c-\rho}{r-\rho}.$$

Evaluation of this last term shows that:

- If $h=0$, then clearly $s^\star=0$ is the minimizing value.
- If $h>0$, then it is easily seen that the infimum over $(0,r-c)$ is attained for

$$s^\star=\min\left\{r-c,\frac{h}{1-h}(c-\rho)\right\}.$$

Substituting into the objective function gives the desired result.

(ii) Lower bound. We distinguish between the cases that (A) $h=0$ and (B) $h>0$.

(A) Take an $\varepsilon>0$, a $\delta\in(0,1)$ and let $\delta^\star$ be defined as in Proposition 3.3(ii). The fact that $h=0$ implies that for arbitrary $\varepsilon'>0$, $v(\delta^\star t)\geq v(t)(1-\varepsilon')$ for t large enough. By arguments similar to those in the upper bound of Theorem 3.6,

$$\begin{aligned}\mathrm{E}\,\mathrm{e}^{\theta A(t)v(t)/t}&\leq\mathrm{e}^{\theta(\rho+\varepsilon)v(t)}+K_\varepsilon\,\mathrm{e}^{\theta rv(t)-v(\delta^\star t)}\\&\leq 2\max\{\mathrm{e}^{\theta(\rho+\varepsilon)v(t)},K_\varepsilon\,\mathrm{e}^{\theta rv(t)-v(t)(1-\varepsilon')}\}\\&\leq 2(1+K_\varepsilon)\max\{\mathrm{e}^{\theta(\rho+\varepsilon)v(t)},\mathrm{e}^{\theta rv(t)-v(t)(1-\varepsilon')}\}.\end{aligned}$$

This implies that

$$I(b)\geq\inf_{t\geq b(r-c)^{-1}}v(t)\sup_\theta\left(\theta\left(\frac{b}{t}+c\right)+\frac{\zeta}{v(t)}-\max\{\theta(\rho+\varepsilon),\theta r-1+\varepsilon'\}\right),$$

with $\zeta:=-\log 2-\log(1+K_\varepsilon)$. The supremum over θ can be explicitly calculated; we obtain the lower bound

$$I(b)\geq\zeta+\inf_{t>b(r-c)^{-1}}v(t)\left(\frac{b/t+c-\rho-\varepsilon}{r-\rho-\varepsilon}(1-\varepsilon')\right).$$

Since $v(b) \to \infty$ as $b \to \infty$,

$$\liminf_{b\to\infty} \frac{I(b)}{v(b)} \geq \liminf_{b\to\infty} \inf_{t>b(r-c)^{-1}} \left(\frac{v(t)}{v(b)}\right)\left(\frac{b/t + c - \rho - \varepsilon}{r - \rho - \varepsilon}(1 - \varepsilon')\right).$$

Now let $\varepsilon \downarrow 0$ and $\varepsilon' \downarrow 0$. Since $t > b(r-c)^{-1}$, $v(t)/v(b) \geq 1 - \eta$ for arbitrary positive η and b large enough. Consequently, $(c-\rho)(r-\rho)^{-1}$ is a lower bound, as desired.

(B) The proof of the lower bound for $h > 0$ is analogous to that in [15]. Assume that $\inf_{s\in(0,r-c)} v(b/s)\bar{J}_{b/s}(s+c)$ is attained for $s = s_b$. (Otherwise, let s_b be such that it reaches a value within ε of the infimum; then let $\varepsilon \downarrow 0$.) Clearly,

$$\liminf_{b\to\infty} \frac{I(b)}{v(b)} = \lim_{b\to\infty} \frac{v(b/s_b)\bar{J}_{b/s_b}(s_b + c)}{v(b)},$$

where the latter limit is along a subsequence. We will denote the subsequence simply by $(s_b)_b$. Then Duffield [15] distinguishes between three cases: (I) there is a closed interval in $(0, r-c)$ in which the $(s_b)_b$ eventually lie, (II) the $(s_b)_b$ converge to 0, and (III) the $(s_b)_b$ converge to ∞. In our setting, the third possibility can clearly be excluded: we have $s_b \in [0, r-c]$ due to peak-rate limitations.

In case (I), we have to use Lemma 3.7(ii): $\bar{J}_t(\cdot)$ converges uniformly on compact subsets of (ρ, r). Also, the regular varying property of $v(\cdot)$ says that $v(b/s)/v(b)$ converges uniformly on closed intervals. Analogously to the proof in [15], this enables us to prove the lower bound immediately.

In case (II), the following reasoning applies. As in [15], it may be shown that $\bar{J}_t(x+c)$ is bounded away from zero as $t \to \infty$, and consequently also $\bar{J}_{b/s_b}(s_b + c)$. As h is positive, $v(b)/v(b/s_b)$ goes to zero. This means that $I(b)/v(b)$ tends to ∞, which contradicts the upper bound. Therefore, a sequence $(s_b)_b$ with limit 0 cannot exist for $h > 0$.

Remark. The proof in [15], for the lower bound in the corresponding $M/G/\infty$ case, does not distinguish between the cases $h = 0$ and $h \in (0, 1)$. In fact, also for the case $h = 0$ it is claimed that there cannot be a subsequence such that $(s_b)_b$ goes to zero. This conclusion relies on the supposition that $s_b \to 0$ implies that $v(b)/v(b/s_b) \to 0$. The latter statement is valid for $h > 0$, but *not* for $h = 0$. Take, for example, $v(b) = \log b$ and $s_b = (\log b)^{-1}$. It is easily verified that $v(\cdot)$ is slowly varying so that $h = 0$. However,

$$\lim_{b\to\infty} \frac{v(b)}{v(b/s_b)} = \lim_{b\to\infty} \frac{\log b}{\log(b \log b)} = 1.$$

In the case of n homogeneous sources, we could circumvent this problem by distinguishing between the cases $h = 0$ and $h \in (0, 1)$.

Nevertheless, we expect the key result (Theorem 4) of [15] to be valid, given the similarity between the model with a fixed number of on–off sources and the one with $M/G/\infty$ input. The proof probably requires more detailed information on the speed of convergence towards the limiting cumulant function found in [37], as in part (A) of our lower bound.

This touches on a crucial distinction between the cases $h = 0$ and $h \in (0, 1)$. Let t_b be the optimizing t for buffer size b, and let us consider b/t_b for b large. When $h = 0$, the proof of the upper bound indicated that $s = 0$ is optimal; in other words, b/t_b approaches 0 for large b, corresponding to t_b being 'superlinear' in b. A similar statement may be derived from the lower bound: t_b is such that $\liminf_{b\to\infty} v(t_b)/v(b) \geq 1$. For $t_b = \alpha b$ (where $\alpha > (r-c)^{-1}$) the

lim inf would give α^{-1}, which is clearly minimized for $\alpha = \infty$. This supports the observation that t_b is superlinear in b.

When $h \in (0, 1)$, the fact that $(s_b)_b$ cannot converge to zero tells us that t_b is approximately linear in b: the time to overflow is proportional to the buffer size. We return to this issue in more detail in the next section.

4. Qualitative insights — reduced load

The results of the previous section may be used to obtain a better understanding of the most likely way for buffer overflow to occur. For Markovian-type sources, very detailed analyses are available. When the on and off periods are mixtures of exponential distributions, it is well understood that the sources must behave according to a different statistical law in order to fill a large buffer. The on and off periods are 'exponentially twisted', such that the on periods are longer and the off periods are shorter than average. References here are the seminal paper of Weiss [45] and recent papers by Mandjes and Ridder [33] and Wischik [47].

For sources with subexponential on periods, the results of the previous section provide the following intuition. During the path to overflow, a source either sends at peak rate for the entire period, or constantly alternates between on and off, and effectively contributes at mean rate. This behavior is nicely reflected in the shape of the cumulant function. Note that this contrasts with the behavior exhibited by Markovian-type sources (as described above), where all sources behave in the same way. A related dichotomy was identified by Anantharam [2], who considered $GI/G/1$ queues. He showed that for exponentially bounded service times, it is multiple long service times and short interarrival times which typically cause overflow, whereas for heavy-tailed service times this is most likely due to just a single extremely long service time.

4.1. Homogeneous sources

In the next two subsections, we develop approximations for the overflow probability $p(B, C)$ based only on knowledge of the distribution of the residual activity period. We consider both the case of homogeneous sources as before, as well as heterogeneous sources.

Conjecture 4.1. (Homogeneous subexponential sources.) *Consider a queue with service rate C fed by n homogeneous on–off sources with $F_{A^\star}(\cdot) \in \mathcal{S} \cap \mathcal{V}$. Then the overflow probability may be approximated as*

$$p(B, C) \approx \max_{K: Kr+(n-K)\rho > C} \mathrm{P}\left(A^\star > \frac{B}{Kr + (n-K)\rho - C}\right)^K,$$

where $K \in \{0, \ldots, n\}$, and $f(B) \approx g(B)$ denotes that the ratio of $\log f(B)$ and $\log g(B)$ tends to 1 for large B. The maximizing value $K^\star$ provides an estimate for the number of sources that send at peak rate during the entire path to overflow.

The approximation may be motivated as follows. Put $K \equiv nk$, and use the scaling $B \equiv nb$ and $C \equiv nc$. Then the conjecture gives

$$\begin{aligned}\frac{1}{n}\log p_n(b, c) &\approx -\min_{k: kr+(1-k)\rho > c} kv\left(\frac{b}{kr + (1-k)\rho - c}\right)\\ &\approx -\min_{k: kr+(1-k)\rho > c} k(kr + (1-k)\rho - c)^{-h} v(b).\end{aligned}$$

The minimum is attained for

$$k^\star = \min\left\{\left(\frac{c-\rho}{r-\rho}\right)\left(\frac{1}{1-h}\right), 1\right\}. \tag{7}$$

Notice that the optimization is equivalent to that in Subsection 3.3 (in the upper bound of Theorem 3.8), with $s = kr+(1-k)\rho-c$; it directly yields the decay rate derived in Theorem 3.8. Thus, the approximation is exact for large n.

For $h = 0$ the fraction of sources that send at peak rate during the path to overflow is $(c-\rho)(r-\rho)^{-1}$, while the remaining fraction $(r-c)(r-\rho)^{-1}$ contribute at mean rate ρ. This gives the aggregate input rate

$$\frac{c-\rho}{r-\rho}r + \frac{r-c}{r-\rho}\rho = c.$$

In other words, if $h = 0$, then the net input rate will only be slightly larger than 0, in agreement with the superlinear time to overflow identified in Subsection 3.3. If $h > 0$, however, then it is easily seen that the net input rate will be strictly positive, thus leading to a time to overflow which is roughly linear in the buffer size. If h is close to 1, then all sources will have long on periods (as $k^\star = 1$). We now illustrate these phenomena through some examples.

Example 1. (*Pareto distribution.*) It is easily verified that if the on periods are Pareto distributed, then $F_{A^*}(\cdot) \in \mathcal{V}$ with $h = 0$; we assume that $v(t) \sim (\alpha - 1)\log t$ for some $\alpha > 1$. In other words, the number of sources sending at peak rate will be such that their peak rates increased by the mean rates of the other sources, just exceed the link rate.

Some calculations show that the decay rate looks like

$$\inf_{t>0} v(t) \sup_{\theta}\left(\theta\left(\frac{b}{t}+c\right) - \frac{\log \mathrm{E}\, e^{\theta A(t)v(t)/t}}{v(t)}\right).$$

With the prior knowledge that t will be large, and Theorem 3.6, the above quantity will approximately be equal to

$$\inf_{t>0} v(t) \sup_{\theta}\left(\theta\left(\frac{b}{t}+c\right) - \max\{\theta\rho, \theta r - 1\}\right) = \inf_{t>0} v(t)\left(\frac{b/t + c - \rho}{r-\rho}\right).$$

Taking the derivative with respect to t yields the first-order condition $b = (c-\rho)t/(\log t -$ solved (for large b) by $t_b = bf(b)$, with $f(\cdot)$ such that $\log(bf(b))/f(b) \to c-\rho$. Note t $f(b)$ is clearly smaller than polynomial, but larger than a constant.

Example 2. (*Lognormal distribution.*) It is easily verified that if the on periods are Lognorma then $F_{A^*}(\cdot) \in \mathcal{V}$ with $h = 0$, as $v(t) \sim 2(\delta \log t)^2$ for a positive parameter δ. In fact, the sam line of reasoning applies as for Pareto on periods. Again, we see that the input rate only slightly exceeds the link rate during the path to overflow, and that the time to overflow is superlinear.

Example 3. (*Weibull distribution.*) A Weibull distribution $\exp[-t^\beta]$ has a $v(\cdot)$ function which is regularly varying of index β. The number of sources that send at peak rate during the path to overflow is given by (7), with $h = \beta$. Notice that, in particular for β close to 1, it may be the case that all sources contribute at peak rate. In any case, the time to overflow will be roughly proportional to the buffer size (as opposed to the case of Pareto distributed or Lognormal on periods), with $t_b k^\star(r-c) \approx b$.

4.2. Heterogeneous sources

In this subsection, we consider the case of heterogeneous sources. First, we focus on a scenario with na sources of type 1 and $n(1-a)$ of type 2, $a \in (0,1)$. Their characteristics (mean, peak, ...) are denoted as usual, but with a subscript to indicate the type of source. Suppose that the type-2 sources are 'smoother' than the type-1 sources, i.e., assume that the type-2 sources have exponential on periods, whereas the type-1 sources have subexponential on periods. Also, assume that $ar_1 + (1-a)\rho_2 > c$. In view of the results of Subsection 4.1, it would seem natural to replace the type-2 sources by their mean rates. We now discuss conditions under which this 'reduced-load approximation' may be justified.

For heterogeneous sources, the analogue of (1) reads:

$$\inf_{t>0} \sup_{\theta} (\theta(b+ct) - a \log \mathrm{E}\, \mathrm{e}^{\theta A_1(t)} - (1-a) \log \mathrm{E}\, \mathrm{e}^{\theta A_2(t)}).$$

Applying the scaling $\theta \to \theta v_1(t)/t$, we obtain

$$\inf_{t>0} v_1(t) \sup_{\theta} \left(\theta \left(\frac{b}{t} + (c - (1-a)\rho_2) \right) - a \frac{\log \mathrm{E} \exp(\theta A_1(t) v_1(t)/t)}{v_1(t)} \right.$$
$$\left. -(1-a) \frac{\log \mathrm{E} \exp(\theta (A_2(t) - (1-a)\rho_2 t) v_1(t)/t)}{v_1(t)} \right).$$

For the reduced-load approximation to apply, the last term in the above expression should vanish for large b. It may be expected that this will be the case if $v_1(t)/\sqrt{t} \to 0$, because of the central limit theorem. However, for A_1 Weibull with shape parameter β larger than $\frac{1}{2}$ this is not the case, and the reduced-load approximation will not hold.

This is in agreement with results of Agrawal *et al.* [1], who consider the case of two sources. They also find that the reduced-load equivalence does not apply for Weibull on periods with $\beta \geq \frac{1}{2}$, on the basis of a different line of reasoning. However, also in their arguments, a crucial role is played by the fact that the central limit theorem does not apply. Recent results by Asmussen *et al.* [5] also confirm — in a different context — the critical value of $\beta = \frac{1}{2}$.

The above observations allow Conjecture 4.1 to be extended to the case of heterogeneous sources.

Conjecture 4.2. (Heterogeneous sources.) *Solve the optimization problem*

$$H(B) = \max_{S} \prod_{i \in S} \mathrm{P}\left(A_i^\star > \frac{B}{\sum_{i \in S} r_i + \sum_{i \notin S} \rho_i - C} \right), \tag{8}$$

with $S \subseteq \{1, \ldots, n\}$ *such that*

$$\sum_{i \in S} r_i + \sum_{i \notin S} \rho_i > C.$$

Assume that the optimizing set $S^\star(B)$ *converges to some set* $S^\star$ *as* $B \to \infty$. *If* $S^\star$ *only consists of sources* i *for which* $v_i(\cdot)$ *is regularly varying of index smaller than* $\frac{1}{2}$, *then* $p(B, C) \approx H(B)$, *where* $f(B) \approx g(B)$ *denotes that the ratio of* $\log f(B)$ *and* $\log g(B)$ *tends to* 1 *for large* B.

The fact that some of the sources show 'peak-rate behavior', while others show 'mean-rate behavior' gives rise to the following conjecture. Denote by Q_S^D the steady-state buffer content in a queue with service rate D fed by the sources $S \subseteq \{1, \ldots, n\}$.

Conjecture 4.3. *If $S^\star$ as defined in Conjecture 4.2 only consists of sources i for which $v_i(\cdot)$ is regularly varying of index smaller than $\frac{1}{2}$, then*

$$\mathrm{P}(Q > B) \sim \mathrm{P}(Q_{S^\star}^{C^\star} > B),$$

where

$$C^\star := C - \sum_{i \notin S^\star} \rho_i,$$

and '$\sim$' denotes that the ratio of both sides tends to 1 for large B.

The above conjecture extends the reduced-load equivalence which has been established in [1] in the special case of two sources. The conjecture has recently been proved in [48] for the general case of multiple heterogeneous sources with regularly varying on periods.

5. Practical implications and conclusions

As mentioned in the introduction, the results of this study may be seen as complementary to those of [32]. There it is shown that in the case of *small* buffers, the tail of the activity period does not have a major effect. The decay rate of the loss probability is completely determined by the *mean* of the on and off periods. It is proved that this insensitivity property still holds when the activity periods are subexponential. Put differently, long-range dependence hardly plays a role in the case of small buffers; we can use a simple exponential on–off model to obtain accurate results. We refer the reader to [20], [22] and [42] for related results illustrating the significant influence of the buffer size on the impact of long-range dependence.

The present paper shows that the opposite holds in case of *large* buffers. In fact, the shape of the residual on-period distribution determines the loss probability. Assuming that the activity periods are exponentially distributed would lead to buffer dimensioning or admission policies which are overly optimistic. As proved in [8], the 'loss curve' $I(b)$ is essentially linear for exponential on periods and large b. The present paper shows that for subexponential on periods the loss curve could well look like, say, $\sqrt{b}$ or $\log b$. In other words, under subexponentiality the tail of the buffer content inherits the essential properties of the tail of the residual activity period.

From a practical perspective, the most relevant scenario is probably the intermediate regime between the two extreme cases discussed above. Unfortunately, we have not been able to obtain results for this regime of *moderate* buffers. In principle, numerical results may be obtained through simulation. The problem is that Monte Carlo techniques are usually slow when a small probability is to be estimated. However, due to the exponentiality in the number of sources n, importance sampling with exponential twisting might be a viable approach.

An interesting topic for further research is also the extension to the case of heterogeneous sources. As indicated in Section 4, this scenario is fundamentally more complicated. The discrepancy between the tails of the on periods of the various types determines whether or not the time to overflow of the 'peak-rate sources' is long enough for the central limit theorem to kick in for the 'mean-rate sources'. A large-deviations analysis of this phenomenon might be possible.

Acknowledgement

We are grateful to A. M. Makowski (University of Maryland, College Park MD, USA) for checking the mathematical details involved in Theorem 3.6.

References

[1] AGRAWAL, R., MAKOWSKI, A. M. AND NAIN, PH. (1999). On a reduced load equivalence for fluid queues under subexponentiality. *Queueing Systems* **33,** 5–41.

[2] ANANTHARAM, V. (1988). How large delays build up in a $GI/G/1$ queue. *Queueing Systems* **5,** 345–368.

[3] ANICK, D., MITRA, D. AND SONDHI, M. M. (1982). Stochastic theory of a data-handling system with multiple sources. *Bell System Tech. J.* **61,** 1871–1894.

[4] ASMUSSEN, S. (1987). *Applied Probability and Queues.* John Wiley, New York.

[5] ASMUSSEN, S., KLÜPPELBERG, C. AND SIGMAN, K. (1999). Sampling at subexponential times, with queueing applications. *Stoch. Proc. Appl.* **79,** 265–286.

[6] BERAN, J., SHERMAN, R., TAQQU, M. S. AND WILLINGER, W. (1995). Long-range dependence in variable-bit-rate video traffic. *IEEE Trans. Commun.* **43,** 1566–1579.

[7] BINGHAM, N. H., GOLDIE, C. AND TEUGELS, J. (1987). *Regular Variation* (Encyclopedia Math. Appl. **27**). Cambridge University Press.

[8] BOTVICH, D. D. AND DUFFIELD, N. G. (1995). Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems* **20,** 293–320.

[9] BOXMA, O. J. (1996). Fluid queues and regular variation. *Perf. Eval.* **27 & 28,** 699–712.

[10] BOXMA, O. J. (1997). Regular variation in a multi-source fluid queue. In *Teletraffic Contributions for the Information Age* (Proc. 15th Int. Teletraffic Congress), eds V. Ramaswami and P. E. Wirth. Elsevier, Amsterdam, pp. 391–402.

[11] BOXMA, O. J. AND DUMAS, V. (1998). Fluid queues with long-tailed activity period distributions. *Comput. Commun.* **21,** 1509–1529.

[12] COHEN, J. W. (1973). Some results on regular variation for distributions in queueing and fluctuation theory. *J. Appl. Prob.* **10,** 343–353.

[13] COURCOUBETIS, C. AND WEBER, R. R. (1996). Buffer overflow asymptotics for a buffer handling many traffic sources. *J. Appl. Prob.* **33,** 886–903.

[14] DUFFIELD, N. G. (1996). Economies of scale in queues with sources having power-law large deviation scalings. *J. Appl. Prob.* **33,** 840–857.

[15] DUFFIELD, N. G. (1998). Queueing at large resources driven by long-tailed $M/G/\infty$-modulated processes. *Queueing Systems* **28,** 245–266.

[16] DUFFIELD, N. G. AND O'CONNELL, N. (1995). Large deviations and overflow probabilities for the general single-server queue, with applications. *Proc. Cambr. Philos. Soc.* **118,** 363–374.

[17] DUMAS, V. AND SIMONIAN, A. (2000). Asymptotic bounds for the fluid queue fed by subexponential on/off sources. *Adv. Appl. Prob.* **32,** 244–255.

[18] ELWALID, A. I. AND MITRA, D. (1991). Analysis and design of rate-based congestion control of high speed networks, I: stochastic fluid models, access regulation. *Queueing Systems* **9,** 29–64.

[19] ERRAMILLI, A., SINGH, R. P. AND PRUTHI, P. (1994). Chaotic maps as model of packet traffic. In *The Fundamental Role of Teletraffic in the Evaluation of Telecommunications Networks* (Proc. 14th Int. Teletraffic Congress), eds J. Labetoulle and J. W. Roberts. Elsevier, Amsterdam, pp. 329–338.

[20] GROSSGLAUSER, M. AND BOLOT, J.-C. (1999). On the relevance of long-range dependence in network traffic. *IEEE/ACM Trans. Networking* **7,** 629–640.

[21] HEATH, D., RESNICK, S. AND SAMORODNITSKY, G. (1997). How system performance is affected by the interplay of averages in a fluid queue with long range dependence induced by heavy tails. *Ann. Appl. Prob.* **9,** 352–375.

[22] HEYMAN, D. AND LAKSHMAN, T. V. (1996). What are the implications of long-range dependence for VBR traffic engineering? *IEEE/ACM Trans. Networking* **4,** 301–317.

[23] JELENKOVIĆ, P. R. AND LAZAR, A. A. (1999). Asymptotic results for multiplexing subexponential on–off processes. *Adv. Appl. Prob.* **31,** 394–421.

[24] KELLA, O. AND WHITT, W. (1992). A storage model with a two-state random environment. *Operat. Res.* **40 (S2),** S257–S262.

[25] KOSTEN, L. (1974). Stochastic theory of a multi-entry buffer, part 1. *Delft Prog. Rept, Ser. F* **1,** 10–18.

[26] KOSTEN, L. (1984). Stochastic theory of data-handling systems with groups of multiple sources. In *Performance of Computer-Communication Systems,* eds H. Rudin and W. Bux. Elsevier, Amsterdam, pp. 321–331.

[27] LELAND, W. E., TAQQU, M. S., WILLINGER, W. AND WILSON, D. V. (1994). On the self-similar nature of Ethernet traffic. *IEEE/ACM Trans. Networking* **2,** 1–15.

[28] LIKHANOV, N. AND MAZUMDAR, R. R. (1998). Cell loss asymptotics in buffers fed with a large number of independent stationary sources. In *Proc. IEEE Infocom '98.* IEEE Computer Society Press, Silver Spring, MD, pp. 339–346.

[29] LIKHANOV, N. AND MAZUMDAR, R. R. (2000). Cell loss asymptotics in buffers fed by heterogeneous long-tailed sources. In *Proc. IEEE Infocom 2000.* IEEE Computer Society Press, Silver Spring, MD, pp. 173–180.

[30] LIKHANOV, N., TSYBAKOV, B. AND GEORGANAS, N. D. (1995). Analysis of an ATM buffer with self-similar ('fractal') input traffic. In *Proc. IEEE Infocom '95*. IEEE Computer Society Press, Silver Spring, MD, pp. 985–992.

[31] LIU, Z., NAIN, PH., TOWSLEY, D. AND ZHANG, Z.-L. (1999). Asymptotic behavior of a multiplexer fed by a long-range dependent process. *J. Appl. Prob.* **36,** 105–118.

[32] MANDJES, M. AND KIM, J.-H. (2001). Large deviations for small buffers: an insensitivity result. To appear in *Queueing Systems*.

[33] MANDJES, M. AND RIDDER, A. (1999). Optimal trajectory to overflow in a queue fed by a large number of sources. *Queueing Systems* **31,** 137–170.

[34] NORROS, I. (1994). A storage model with self-similar input. *Queueing Systems* **16,** 387–396.

[35] NORROS, I. (1995). On the use of fractional Brownian motion in the theory of connectionless networks. *IEEE J. Sel. Areas Commun.* **13,** 953–962.

[36] PAKES, A. G. (1975). On the tail of waiting time distributions. *J. Appl. Prob.* **12,** 555–564.

[37] PARULEKAR, M. AND MAKOWSKI, A. M. (1997). Tail probabilities for a multiplexer driven by $M/G/\infty$ input processes (I): preliminary asymptotics. *Queueing Systems* **27,** 271–296.

[38] PAXSON, V. AND FLOYD, S. (1995). Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Trans. Networking* **3,** 226–244.

[39] RESNICK, S. AND SAMORODNITSKY, G. (1999). Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues. *Queueing Systems* **33,** 43–71.

[40] RESNICK, S. AND SAMORODNITSKY, G. (1999). Steady state distribution of the buffer content for $M/G/\infty$ input fluid queues. Tech. Rept TR1242, Cornell University.

[41] ROCKAFELLAR, R. T. (1970). *Convex Analysis*. Princeton University Press.

[42] RYU, B. K. AND ELWALID, A. I. (1996). The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities. *Comput. Commun. Rev.* **26,** 3–14.

[43] SIMONIAN, A. AND GUIBERT, J. (1995). Large deviations approximation for fluid queues fed by a large number of on/off sources. *IEEE J. Sel. Areas Commun.* **13,** 1017–1027.

[44] STERN, T. E. AND ELWALID, A. I. (1991). Analysis of separable Markov-modulated rate models for information-handling systems. *Adv. Appl. Prob.* **23,** 105–139.

[45] WEISS, A. (1986). A new technique of analyzing large traffic systems. *Adv. Appl. Prob.* **18,** 506–532.

[46] WILLINGER, W., TAQQU, M. S., SHERMAN, R. AND WILSON, D. V. (1997). Self-similarity through high variability: statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Trans. Networking* **5,** 71–86.

[47] WISCHIK, D. J. (2001). Sample path large deviations for queues with many inputs. To appear in *Queueing Systems*.

[48] ZWART, A. P., BORST, S. C. AND MANDJES, M. (2000). Exact queueing asymptotics for multiple long-tailed on–off sources. Tech. Rept SPOR 2000-14, Eindhoven University of Technology.