Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

Overflow behavior in queues with many long-tailed inputs

M.R.H. Mandjes, S.C. Borst

Probability, Networks and Algorithms (PNA)

PNA-R9911 November 30, 1999

Report PNA-R9911
ISSN 1386-3711

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

CWI is the National Research Institute for Mathematics and Computer Science. CWI is part of the Stichting Mathematisch Centrum (SMC), the Dutch foundation for promotion of mathematics and computer science and their applications.
SMC is sponsored by the Netherlands Organization for Scientific Research (NWO). CWI is a member of ERCIM, the European Research Consortium for Informatics and Mathematics.

Copyright © Stichting Mathematisch Centrum
P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

Overflow Behavior in Queues with Many Long-Tailed Inputs

Michel Mandjes*, Sem Borst*,†,‡

*Bell Laboratories, Lucent Technologies
Murray Hill, NJ 07974, USA

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

‡Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

ABSTRACT

We consider a fluid queue fed by a superposition of n homogeneous on-off sources with generally distributed on- and off-periods. We scale buffer space B and link rate C by n, such 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 situation in which also b is large. We explicitly compute asymptotics for the case in which the on-periods have a subexponential distribution, e.g., Pareto, Lognormal, or Weibull.

We provide a detailed interpretation of our results. Crucial is the shape of the function $v(t) := -\log \mathbb{P}(A^\star > t)$ for large t, $A^\star$ being the *residual* on-period. If $v(\cdot)$ is slowly varying (e.g., Pareto, Lognormal), then, during the trajectory 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 essentially 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 other sources contribute at mean rate. These observations lead to straightforward approximations for the overflow probability. These approximations can be extended to the case of heterogeneous sources; the results provide further insight into the so-called reduced-load approximation.

1991 Mathematics Subject Classification: 60K25 (primary), 68M20, 90B12, 90B22 (secondary).

Keywords and Phrases: buffer overflow, large-deviations asymptotics, long-tailed on-periods, long-range dependence, on-off sources, queueing theory, reduced-load approximation, regular variation, subexponentiality.

Note: Work of the second author carried out in part under the project PNA2.1 "Communication and Computer Networks".

1 Introduction

Measurements have indicated that network traffic exhibits burstiness on a wide range of time scales [25]. This conclusion was drawn after thorough examination of traffic streams in a variety of packet-based networks, see e.g. [6, 35]. The discovery of the presence of *long-range dependence* had an important impact on teletraffic research. Where one used to rely on short-range dependent models, recently increasing insight has been gained into traffic models which exhibit burstiness on a wider range of time scales.

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 can be distinguished. (i) Erramilli *et al.* [19] advocate the application of *chaotic maps*. (ii) Norros [31, 32] proposes the use of queues fed by *fractional Brownian motion*. (iii) Willinger *et al.* [41] use the superposition of on-off fluid sources with *long-tailed* activity periods. The present paper will follow the third approach.

An on-off fluid source alternates between activity periods (commonly called 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 often focuses on the probability of the buffer content exceeding level B. In a series of papers, Kosten, e.g. [23, 24], and Mitra *et al.*, e.g. [3, 18], examine the case in which 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 their models the overflow probability decays essentially exponentially in B. In this paper we will depart from the assumption that the on- and off-periods have exponentially bounded tails, but rather allow 'long-tailed' distributions, like Pareto, Lognormal, and Weibull. We expected that long-tailed bursts would give rise to overflow probabilities that decay slower than exponentially in the buffer level B. This expectation was based on a couple of results in the literature.

Literature

The early literature on long-tailed queues goes back to the seventies. The most important contribution was by Cohen [12] and Pakes [34]. They consider GI/G/1 queues in which the residual service times have 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 the same shape as the residual service time.

From 1995 the focus shifted to on-off fluid sources. In Boxma [9] and Jelenković and Lazar [21], an analysis is given of a queue fed by a single source. Applying the result for the GI/G/1 queue, they managed to find the large-buffer asymptotics. It appears that this tail is mainly

determined by the distribution of the residual burst duration – provided that this random variable has a subexponential distribution. This result plays a crucial role in our analysis.
The results for the case of *multiple* on-off sources are rather limited. Three types of results are worth mentioning here. (i) In [9, 11] and [21] large-buffer asymptotics are derived for the case in which the peak rate of a single source is already larger than the link rate. (ii) Dumas and Simonian [17] find asymptotic upper and lower bounds on the overflow probability. These bounds are in general not sharp for all subexponential on-periods; a tightness property can only be derived for the subclass of regularly varying burst durations. (iii) Agrawal, Makowski, and Nain [1] consider a queue fed by two on-off sources. Under particular conditions this queue is proven to be asymptotically equivalent to the queue fed by one of these sources, emptied at rate C subtracted by the mean rate of the other source. This *reduced-load approximation* was also established by [10, 21] for the situation of one heavy-tailed source and a number of exponential sources. Here it was assumed that the on-period of the heavy-tailed source had a burst duration which is of (intermediate) regular variation; at the same time it is required that the peak rate of the heavy-tailed source, increased by the mean rates of the other sources exceeds the link rate C.
An interesting related class of models is that of M/G/∞ input. Sessions arrive according to a Poisson process, and remain in the system during a generally distributed time, during which they generate traffic at a fixed rate. Parulekar and Makowski [33] 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 $n\lambda$, $C \equiv nc$, 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 are given in Likhanov *et al.* [27] and Liu *et al.* [28].

Contribution

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 buffer overflow probability decays exponentially in the scaling parameter n, with decay rate $I(b)$ as 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 two significant 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 first characterizing the moment generating function of the traffic generated by a single source in an interval of length t, in a specific scaling. This scaling was proposed by Parulekar and Makowski [33] for the M/G/∞ case, but also applies for our model. Then we exploit this characterization to establish the large-b asymptotics for $I(b)$. As

pointed out above, all large-buffer asymptotics obtained before 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 have been too optimistic in the case of large buffers. This result can be seen as complementary to the conclusions of Mandjes and Kim [29]; there it was shown that for small buffers the long-range dependent behavior of the sources does not significantly affect the loss statistics. We refer to Heyman and Lakshman [20] and Ryu and Elwalid [37] for related results indicating the critical role of the buffer size in assessing the impact of long-range dependence.

- We contribute to the understanding of *the way overflow occurs*. Clearly, to fill a large buffer, there is always the trade-off between the *intensity* of the deviant behavior (to what extent does the input rate exceed the link rate?) and the *duration* of that deviant behavior. For on-off sources with exponentially bounded burst times, it was already found that sources alternate between on and off during the path to overflow, but with longer on-periods and shorter off-periods; all sources essentially behave in the same way [2, 30, 40]. In contrast, if the burst durations are subexponential, then sources contribute either at their mean rate or their peak rate. Put differently: some sources stay in the on-state during the entire trajectory, 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, also for the situation where there is a heterogeneous mix of sources. For the class of regularly varying sources, this gives the reduced-load approximation, which is in agreement with the bounds of Dumas and Simonian [17].

Organization

Section 2 describes the model, introduces notation, and gives some basic definitions and assumptions. Section 3 presents the analysis. We prove 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 our model and the required assumptions. The second subsection briefly reviews the class of subexponential distribution functions 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. This 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. Define also

$$A(t) := \text{Traffic generated by a single source in steady state in time interval } [0,t].$$

Later in our analysis we need the following assumption.

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

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

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

and the fraction spent in the off-state is its 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

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

We are interested in the probability of the buffer content exceeding 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 [40] and has proven to be very powerful, see e.g. the contributions by Botvich and Duffield [8], Courcoubetis and Weber [13], and Simonian and Guibert [38]. We assume that the system is stable and non-trivial:

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

In the scaled model we define

$$p_n(b,c) := \text{steady-state probability that the buffer content exceeds level } nb.$$

In particular we will analyze its 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 this limit exists.

2.2 Subexponentiality

This subsection gives the definition of subexponential distribution functions and states the asymptotic result for a queue fed by a single source with subexponential on-periods. More details can be found in the appendices of [11]. Throughout, we denote by $F_X(\cdot)$ the distribution function of the random variable X, with specifically $F_{A^\star}(\cdot)$ indicating the distribution function of the residual activity period.

Definition 2.2 [Subexponential distribution]
Suppose

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

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

$$v_X(t) := -\log \mathbb{P}(X > t).$$

In Section 3 we will show that the shape of this function determines the large-buffer asymptotics, with $X = A^\star$.

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

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

for all $y > 0$. If $v_X(\cdot)$ is regularly varying of index $h \in [0, 1)$, 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 Bingham, Goldie, and Teugels [7, Section 1.4]. Unfortunately, the exact relation between the classes $\mathcal{S}$ and $\mathcal{V}$ is not clear. However, the most important long-tailed distributions (like Pareto, Lognormal, and Weibull) are in both of them.

The following theorem is an extension of the results for the GI/G/1 queue by Cohen [12] and Pakes [34], and can be found in Boxma [9] and Jelenković and Lazar [21]. This result for a queue fed by a single source will be one of the main building blocks in our analysis of the queue fed by n sources.

Theorem 2.4 [Single source]
Consider a single source feeding into a queue with service rate c, and let Q denote the steady-state queue length. If $F_{A^\star}(\cdot) \in \mathcal{S}$ and $\rho < c < r$, then

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

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

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 space 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 played by 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*. This cumulant function is needed in the variational problem mentioned above. We show that under a certain scaling this cumulant function is piecewise linear for on-off sources with subexponential on-periods. The choice of this particular scaling is due to Duffield [15] and Parulekar and Makowski [33].
The third subsection contains the main theorem: the asymptotics (for large b) of the decay rate. We combine the variational problem of Section 3.1 and the cumulant function of Section 3.2. Here we follow the approach that Duffield [15] used for the M/G/∞ case.

3.1 Decay rate for general buffer size

In this subsection, we focus on the evaluation of the decay rate for general buffer level b. The theorem which we will use in Section 3.3 is a variant of the key theorem of Likhanov and Mazumdar [26], which is stated in Theorem 3.2. Their result was phrased in the setting of slotted time; in Mandjes and Kim [29] it was extended to continuous time. The latter version is formulated in Theorem 3.2 below, and requires the following assumption.

Assumption 3.1
Define

$$J_t(x) := \sup_\theta \left(\theta x - \log \mathbb{E}e^{\theta A(t)}\right).$$

Assume

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

(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,

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

For the proof of Theorem 3.2 we refer to Mandjes and Kim [29]. 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 [26]. In earlier versions of Theorem 3.2, e.g., the one in Botvich and Duffield [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

$$\log \mathbb{E}e^{\theta A(t)} < c\theta t.$$

It is not hard to show that this condition is not fulfilled for on-off sources with heavy-tailed on-periods, see Mandjes and Kim [29]. The (weaker) requirement Assumption 3.1.(i) is satisfied under the non-restrictive condition that $\mathbb{E}A^{1+\zeta}$ is finite for some positive ζ, as we postulated in Assumption 2.1. We prove this in the next proposition.

Proposition 3.3
Consider an on-off source with on-periods A.

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

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

$$\mathbb{P}\left(\frac{A(t)}{t} > \rho + \epsilon\right) \leq K_\epsilon \cdot \mathbb{P}(A^\star > \delta^\star t) \quad \textit{with} \quad \delta^\star := \frac{(1-\delta)\epsilon}{r - (\rho + \delta\epsilon)},$$

$\delta \in (0,1)$.

Proof

We first prove the second statement.

(ii) We may write

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

where Q is defined as the steady-state queue length in case the source feeds into a queue which is emptied at rate $\rho + \delta\epsilon$. Invoking Theorem 2.4, we see that there is a K_ϵ such that for t large enough the stated holds.

(i) Proposition 3.1 of Likhanov and Mazumdar [26] states that Assumption 3.1.(i) is satisfied if for all $\epsilon \in (0, r-\rho)$ there are positive K and α such that for t large enough

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

As above,

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

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

To see that the stated also holds in case $F_{A^\star}(\cdot) \notin \mathcal{S}$, (in which case Theorem 2.4 cannot be invoked), we may use a result of Kella and Whitt [22] to relate the queue length 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 $\mathbb{E}W^\zeta < \infty$ if $\mathbb{E}A^{1+\zeta} < \infty$ 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

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

for an increasing positive function $w(\cdot)$. Compared to the variational problem of (1), the optimum is attained at the same $t^\star$ and a value of $\theta^\star$ that is $t^\star/w(t^\star)$ 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 in the buffer size. For the model with M/G/∞ input, it was proposed in [33] to use $w(t) = -\log \mathbb{P}(D^\star > t)$, where $D^\star$ is the residual session duration. In the sequel we use the scaling

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

3.2 The cumulant function

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

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

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 some intuitive background 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

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

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

Lemma 3.5

For all θ, with $h \in [0, 1)$,

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

Here x^h with $h = 0$ is defined by 1 for $x > 0$, and by 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

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

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 $\theta \leq \theta_0$, and smaller than or equal to $\theta r - 1$ for $\theta \geq \theta_0$.
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 $F_{A^\star}(\cdot) \in \mathcal{S} \cap \mathcal{V}$ and $\theta \geq 0$,

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

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

Proof
The proof consists of a lower bound and an upper bound.

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

  Also, a 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 $\theta \geq 0$ and all $t > 0$:

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

  where p denotes the steady-state on-probability.

  Both lower bounds together yield the desired result:

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

- **Upper bound** First take $\theta \geq 0$. Choose a $k \in \mathbb{N}$ and $\epsilon := \epsilon_k = (r-\rho)/k$.

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

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

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

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

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

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

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

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

3.3 Large-buffer asymptotics of the decay rate

In this subsection we will combine the large-deviations results for general buffer size 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 the proof of Duffield [15] for the M/G/∞ 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 $x \in (\rho, r)$

$$\lim_{t\to\infty} \bar{J}_t(x) = \frac{x-\rho}{r-\rho}.$$

(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 also the Legendre-Fenchel transforms $f_t^\star(x) := \sup_\theta(\theta x - f_t(\theta))$ converge to $f^\star(x) := \sup_\theta(\theta x - f(\theta))$ on the interior of the effective domain of $f^\star$.
Therefore, for $x \in (\rho, r)$,

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

From the fact that $\mathbb{E}A(t) = \rho t$, in conjunction with $x \in (\rho, r)$, it easily follows that the above supremum needs to be taken over positive θ only. Recalling Theorem 3.6 and noticing 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 can be found in Rockafellar [36, Theorem 10.8], the following property holds. If finite convex functions f_t converge pointwise to a finite convex function f on a certain domain, then this convergence is uniform on compact subsets of the domain.

So we are left 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

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

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

We are now in a position to prove our main theorem. It states that for on-off sources with $F_{A^\star}(\cdot) \in \mathcal{V}\cap\mathcal{S}$, the $I(b)$ curve 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 $F_{A^\star}(\cdot) \in \mathcal{S}\cap\mathcal{V}$*:*

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

Proof

The proof consists of an upper bound and a lower bound.

- **Upper bound** The proof of the upper bound is parallel to that given by Duffield [15] for the M/G/∞ 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 easy; the intuition is that t represents the epoch of overflow starting in an empty system; this must obviously be later than $b\,(r-c)^{-1}$.

Now substituting $b/t =: s$, 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 $s \in (0, r-c)$,

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

Explicit evaluation of this last term shows

- 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.

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

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

  $$\begin{aligned}
  \mathbb{E}e^{\theta A(t)v(t)/t} &\leq e^{\theta(\rho+\epsilon)v(t)} + K_\epsilon e^{\theta r v(t) - v(\delta^\star t)} \\
  &\leq 2\max\left\{e^{\theta(\rho+\epsilon)v(t)}, K_\epsilon e^{\theta r v(t) - v(t)(1-\epsilon')}\right\} \\
  &\leq 2(1+K_\epsilon)\cdot\max\left\{e^{\theta(\rho+\epsilon)v(t)}, e^{\theta r v(t) - v(t)(1-\epsilon')}\right\}.
  \end{aligned}$$

  This implies that

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

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

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

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

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

Now let $\epsilon \downarrow 0$ and $\epsilon' \downarrow 0$. As $t > b(r-c)^{-1}$, $v(t)/v(b) \geq 1-\eta$ 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 Duffield [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 $\epsilon \downarrow 0$ at the end of the procedure.) 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 this 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, (iii) the $(s_b)_b$ converge to ∞. In our setting, clearly the third possibility can 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 Duffield [15], it can 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 Duffield [15], for the lower bound in the corresponding M/G/∞ case, does not distinguish between the cases $h = 0$ and $h \in (0,1)$. In fact, also for $h = 0$ it is claimed that there cannot be a subsequence such that $(s_b)_b$ goes to zero. To draw this conclusion it is used that $s_b \to 0$ implies $v(b)/v(b/s_b) \to 0$. This statement is valid for $h > 0$, but *not* for $h = 0$. Take for instance $v(b) = \log b$ and $s_b = (\log b)^{-1}$. It is easily verified that $v(\cdot)$ is slowly varying so $h = 0$. However,

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

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

Nonetheless, we expect that the key statement (Theorem 4) of [15] is valid, given the similarity between the model with a fixed number of on-off sources and the one with M/G/∞ input. Its proof probably requires more information on the speed of convergence towards the limiting cumulant function found by [33], like 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. In the former case, i.e., $h = 0$, the proof of the upper bound gave 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 can be derived from the lower bound: t_b is such that

$$\liminf_{b\to\infty} \frac{v(t_b)}{v(b)} \frac{b}{t_b} \geq 1.$$

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.

In the latter case, i.e., $h \in (0, 1)$, the fact that $(s_b)_b$ cannot converge to zero says that t_b will be essentially linear in b: the time to overflow will be proportional to the buffer size. We will return to this issue in more detail in the next section.

4 Qualitative insights – reduced load

The theoretical results of the previous section enable us to obtain a better understanding of the most likely way for buffer overflow to occur. For Markovian-type sources, very detailed analyses are available. In case 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-times are exponentially twisted, such that the on-periods are longer and the off-periods are shorter. References here are the seminal paper of Weiss [40], and recent papers by Mandjes and Ridder [30] and Wischik [42].

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 the entire time interval, or constantly alternates between on and off, and effectively contributes at mean rate. This 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 essentially 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 one extremely long service time.

4.1 Homogeneous sources

In the next two subsections we develop approximations for the overflow probability $p(B,C)$ only based on knowledge of the distribution of the residual burst durations. We will do that for both the setting with homogeneous sources, as was assumed before, as well as heterogeneous sources.

Approximation 4.1 [Homogeneous subexponential sources]
Consider a queue with buffer B and rate C fed by n homogeneous on-off sources with $F_{A^\star}(\cdot) \in \mathcal{S} \cap \mathcal{V}$. An approximation for the loss probability is

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

where the value of K is restricted to $\{0,\ldots,n\}$, and $\approx$ indicates that the ratio of both sides tends to a fixed constant for large B. The maximizing value $K^\star$ provides an estimate for the number of sources that send at peak rate during the entire time interval to overflow.

The justification for this approximation is as follows. Put $K \equiv nk$, and use the scaling $B \equiv nb$ and $C \equiv nc$. Then

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

The minimum is attained at

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

Notice that the optimization is equivalent to that in Section 3.3 (in the upper bound of Theorem 3.8), where $s = kr+(1-k)\rho-c$. It directly leads to the decay rate derived in Theorem 3.8. For $h=0$ the fraction of sources that send at peak rate during the path to overflow is $(c-\rho)(r-\rho)^{-1}$, whereas the remaining fraction $(r-c)(r-\rho)^{-1}$ contribute at mean rate ρ. This gives aggregate input rate

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

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 Section 3.3. If $h>0$, however, then it is easily seen that the net input rate will be positive, thus leading to a time to overflow which is essentially linear in the buffer size. If h is close to 1, then all sources will have long bursts (as $k^\star = 1$). We now illustrate these phenomena through some examples.

Example 1 (Pareto)

It is easily checked that if the on-periods are Pareto distributed, then $F_{A^\star}(\cdot) \in \mathcal{V}$ 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 exceeds the link rate.

Here we do some rough calculations on the decay rate, and the optimizing arguments in the inf sup problem, as they considerably contribute to the understanding of the way overflow occurs. These calculations show that the decay rate looks like

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

With the prior knowledge that t will be large, and Theorem 3.6, this 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 w.r.t. t yields the first-order condition

$$\frac{b}{t} + c - \rho = \frac{b \log t}{t},$$

solved (b large) for $t_b = bf(b)$, with $f(\cdot)$ such that $\log(bf(b))/f(b) \to 1$. Thus $f(b)$ is clearly smaller than polynomial, but larger than a constant.

Example 2 (Lognormal)

It is easily checked that for A Lognormal we have that $F_{A^\star}(\cdot) \in \mathcal{V}$ with $h = 0$, as $v(t) \sim 2(\delta \log t)^2$ for a positive parameter δ. In fact, the same line of reasoning applies as for Pareto on-periods. Again, we see that the input rate is only slightly larger than the link rate, and the time to overflow is superlinear.

Example 3 (Weibull)

A Weibull distribution with A having cumulative distribution function $\exp[-t^\beta]$ has a $v(\cdot)$ function which is regularly varying of index β. The number of sources that show 'peak-rate behavior' is given by (7), with $h = \beta$. Notice that, in particular for β close to 1, it can be the case that during the most likely path to overflow all sources send 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 or Lognormal on-periods), with $t_b k^\star(r - c) \approx b$.

4.2 Heterogeneous sources

In this subsection we discuss the case of heterogeneous sources. First we consider the situation in which there are 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 for instance that the type-2 sources have exponential on-periods, whereas the type-2 sources have subexponential on-periods. Also assume that $ar_1 + (1-a)\rho_2 > c$. Under these conditions it is tempting to replace the type-2 sources by their mean rate, in view of the results of Section 4.1. Here we will discuss the conditions under which this 'reduced-load approach' is justified.

For heterogeneous sources, the analogue of (1) reads

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

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

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

For reduced load to apply, the last term in the above expression should vanish for large b. It can 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 being Weibull with shape parameter β larger than $\frac{1}{2}$ this is not the case, and reduced load will not hold.

This is in agreement with the results in Agrawal, Makowski and Nain [1], who consider the case of two sources. They also found that reduced load does not apply for Weibull sources with $\beta \geq \frac{1}{2}$, on the basis of a different line of reasoning. However, also in their arguments, a crucial role was played by the fact that the central limit theorem does not apply. Recent results by Asmussen, Klüppelberg, and Sigman [5] also confirm – in a different context – the critical value of $\beta = \frac{1}{2}$

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

Approximation 4.2 [Heterogeneous sources]
Solve the optimization problem

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

where $S \subseteq \{1, \ldots, n\}$ *is 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* $S^\star$ *for* $B \to \infty$. *If* $S^\star$ *only consists of sources* i *for which* $v_i(\cdot)$ *is regularly varying of index smaller than* $\frac{1}{2}$, *then* $p(B,C) \approx G(B)$, *where* $\approx$ *denotes that the ratio of the two quantities tends to a fixed constant for large* B.

The fact that some of the sources contribute at their mean rate and others at their peak rate gives rise to the following conjecture. Let Q_S^D be the steady-state buffer content of the queue fed by the sources $S \subseteq \{1, \ldots, n\}$, emptied at rate D.

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

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

where $C^\star := C - \sum_{i \notin S^\star} \rho_i$, and '$\sim$' denotes that the ratio of both sides tends to 1 when $B \to \infty$,

5 Practical implications and conclusions

As mentioned in the introduction, this study can be regarded as complementary to Mandjes and Kim [29]. There we showed that in the case of *small* buffers, the tail of the activity period did not play a role at all: the decay rate of the loss probability is completely determined by the *mean* of the on- and off-periods. We proved that this insensitivity property still holds in case the activity periods are subexponential. Put differently: the phenomenon of long-range dependence plays hardly a role in case the buffer is small. We can use the simple exponential on-off model to get accurate results. We refer to Heyman and Lakshman [20] and Ryu and Elwalid [37] for related results illustrating the critical role of the buffer size in determining the impact of long-range dependence.

The present paper shows that the opposite conclusion holds in case of *large* buffers. In fact, the shape of the distribution of the residual burst duration determines the loss behavior. Using the assumption that the activity periods are exponentially distributed could lead to buffer dimensioning or admission policies that are too optimistic. As shown in [8], the 'loss curve' $I(b)$ is essentially linear for exponential bursts and large b; in this paper however we showed that for subexponential bursts this curve could well look like, say, $\sqrt{b}$ or $\log b$. In other words: under subexponentiality the tail of the queue inherits the essential properties of the tail of the residual activity periods.

In practice however, the most relevant scenario is probably between the two extremes described above. Unfortunately, we have not been able to obtain results for this regime of *moderate* buffers. One way to obtain numerical values is to resort to simulation; the problem of course is that these Monte Carlo techniques are usually slow as 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.

Another interesting subject for future research is the extension to heterogeneous sources. As mentioned in Section 4, this situation is fundamentally more complex: the discrepancy between

the tails of the on-periods determines whether the time to overflow of the 'peak-rate sources' is long enough to make the central limit theorem kicking in for the 'mean-rate sources'. A large-deviations analysis of this phenomenon might be possible.

Acknowledgment

A. Makowski (University of Maryland, College Park MD, USA) is gratefully acknowledged for checking the mathematical details involving Theorem 3.6.

References

[1] R. Agrawal, A. Makowski, and P. Nain. On a reduced load equivalence for fluid queues under subexponentiality. To appear in: *Queueing Systems*, 1999.

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

[3] D. Anick, D. Mitra, and M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *The Bell System Technical Journal*, 61: 1871–1894, 1982.

[4] S. Asmussen. *Applied Probability and Queues.* Wiley, New York, 1987.

[5] S. Asmussen, C. Klüppelberg, and K. Sigman. Sampling at subexponential times, with queueing applications. *Stochastic Processes and their Applications*, 79: 265–286, 1999.

[6] J. Beran, R. Sherman, M. Taqqu, and W. Willinger. Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications*, 43: 1566–1579, 1995.

[7] N. Bingham, C. Goldie, and J. Teugels. *Regular Variation.* Encyclopedia of Mathematics and its Applications, Vol. 27. Cambridge University Press, Cambridge, 1987.

[8] D. Botvich and N. Duffield. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems*, 20: 293–320, 1995.

[9] O. Boxma. Fluid queues and regular variation. *Performance Evaluation*, 27 & 28: 699–712, 1996.

[10] O. Boxma. Regular variation in a multi-source fluid queue. *Proceedings ITC 15*, 391–402, 1997.

[11] O. Boxma and V. Dumas. Fluid queues with long-tailed activity period distributions. *Computer Communications*, 21: 1509–1529, 1998.

[12] J. Cohen. Some results on regular variation for distributions in queueing and fluctuation theory. *Journal of Applied Probability*, 10: 343–353, 1973.

[13] C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a buffer handling many traffic sources. *Journal of Applied Probability*, 33: 886–903, 1996.

[14] N. Duffield. Economies of scale in queues with sources having power-law large deviation scalings. *Journal of Applied Probability*, 33: 840–857, 1996.

[15] N. Duffield. Queueing at large resources driven by long-tailed M/G/∞-modulated processes. *Queueing Systems*, 28: 245–266, 1998.

[16] N. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single server queue, with applications. *Proceedings of the Cambridge Philosophical Society*, 118: 363–374, 1995.

[17] V. Dumas and A. Simonian. Asymptotic bounds for the fluid queue fed by subexponential on-off sources. *Preprint.*

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

[19] A. Erramilli, R. Singh, and P. Pruthi. Chaotic maps as model of packet traffic. *Proceedings ITC 14*, 329–338, 1994.

[20] D. Heyman and T. Lakshman. What are the implications of long-range dependence for VBR traffic engineering? *IEEE/ACM Transactions on Networking*, 4: 301–317, 1996.

[21] P. Jelenković and A. Lazar. Asymptotic results for multiplexing subexponential on-off processes. *Advances in Applied Probability*, 31: 394-421, 1999.

[22] O. Kella and W. Whitt. A storage model with a two-state random environment. *Operations Research*, 40(S2): S257-S262, 1992.

[23] L. Kosten. Stochastic theory of a multi-entry buffer, part 1. *Delft Progress Report, Series F*, 1: 10–18, 1974.

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

[25] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Transactions on Networking*, 2: 1–15, 1994.

[26] N. Likhanov and R. Mazumdar. Cell loss asymptotics in buffers fed with a large number of independent stationary sources. *Proceedings IEEE Infocom*, 339–346, 1998.

[27] N. Likhanov, B. Tsybakov, and N. Georganas. Analysis of an ATM buffer with self-similar ('fractal') input traffic. *Proceedings IEEE Infocom*, 985–992, 1995.

[28] Z. Liu, P. Nain, D. Towsley, and Z.-L. Zhang. Asymptotic behavior of a multiplexer fed by a long-range dependent process. *Journal of Applied Probability*, 36: 105–118, 1999.

[29] M. Mandjes and J.-H. Kim. Large deviations for small buffers: an insensitivity result. Submitted to: *Queueing Systems.*

[30] M. Mandjes and A. Ridder. Optimal trajectory to overflow in a queue fed by a large number of sources. *Queueing Systems*, 31: 137–170, 1999.

[31] I. Norros. A storage model with self-similar input. *Queueing Systems*, 16: 387–396, 1994.

[32] I. Norros. On the use of fractional Brownian motion in the theory of connectionless networks. *IEEE Journal on Selected Areas in Communications*, 13: 953–962, 1995.

[33] M. Parulekar and A. Makowski. Tail probabilities for a multiplexer driven by M/G/∞ input processes (I): preliminary asymptotics. *Queueing Systems*, 27: 271–296, 1997.

[34] A. Pakes. On the tail of waiting time distributions. *Journal of Applied Probability*, 12: 555–564, 1975.

[35] V. Paxson and S. Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3: 226–244, 1995.

[36] R. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, 1970.

[37] B. Ryu and A. Elwalid. The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities. *Computer Communication Review*, 26: 3–14, 1996.

[38] A. Simonian and J. Guibert. Large deviations approximation for fluid queues fed by a large number of on/off sources. *IEEE Journal of Selected Areas in Communications*, 13: 1017–1027, 1995.

[39] T. Stern and A. Elwalid. Analysis of separable Markov-modulated rate models for information-handling systems. *Advances in Applied Probability*, 23: 105–139, 1991.

[40] A. Weiss. A new technique of analyzing large traffic systems. *Advances in Applied Probability*, 18: 506–532, 1986.

[41] W. Willinger, M. Taqqu, R. Sherman, and D. Wilson. Self-similarity through high variability: statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking*, 5: 71–86, 1997.

[42] D. Wischik. Sample path large deviations for queues with many inputs. To appear in: *Queueing Systems*, 1999.