ELSEVIER

Int. J. Electron. Commun. (AEÜ) **58** (2004): 101–117
http://www.elsevier-deutschland.de/aeue

International Journal AEÜ of **Electronics** and **Communications**

The Shape of the Loss Curve and the Impact of Long-Range Dependence on Network Performance

Michel Mandjes and Nam Kyoo Boots

Abstract: Empirical studies showed that many types of network traffic exhibit *long-range dependence* (LRD), i.e., burstiness on a wide variety of time-scales. Given that traffic streams are indeed endowed with LRD properties, a next question is: what is their impact on network performance?

To assess this issue, we consider a generic source model: traffic generated by an individual user is modeled as a fluid on/off pattern with generally distributed on- and off-times; LRD traffic is obtained by choosing the on-times heavy-tailed. We focus on an aggregation of many i.i.d. sources, say n, multiplexed on a FIFO queue, with the queueing resources scaled accordingly. Large deviations analysis says that the (steady-state) overflow probability decays exponentially in n; we call the corresponding decay rate, as a function of the buffer size B, the *loss curve*.

To get insight into the influence of the distribution of the on- and off-times, we list the most significant properties of the loss curve. Strikingly, for small B, the decay rate depends on the distributions *only through their means*. For large B there is no such insensitivity property. In case of heavy-tailed on-times, the decay of the loss probability in the buffer size is slower than exponential; this is in stark contrast with light-tailed on-times, in which case this decay is at least exponential.

To assess the sensitivity of the performance metrics to the probabilistic properties of the input, we compute the loss curve for a number of representative examples (voice, video, file transfer, web browsing, etc.), with realistic distributions and parameters.

Our conclusions on the impact of LRD on the performance can be summarized as follows: (1) If the maximally tolerable delay is relatively small, there is hardly any difference between heavy-tailed and light-tailed inputs; this gives a theoretical handle on observations that appeared in the literature. Only for very delay tolerant applications the above-mentioned large B results kick in. (2) The level of aggregation is a significant factor. If the ratio between the link rate and the peak rate of a single source is high, a high utilization can be achieved, while at the same time the delay requirements are met; this holds even if the delay requirements are stringent.

Keywords: Packet networking, Long-range dependence, Queueing theory, Large deviations asymptotics, Buffer overflow, Heavy-tailed distributions

Received October 25, 2002. Revised June 10, 2003.

M. Mandjes, CWI, Advanced Communication Networks, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands.
Part of this work was done while at Bell Laboratories/Lucent Technologies, Murray Hill, NJ 07974-0636, USA.
N.K. Boots, ABN AMRO Bank, Gustav Mahlerlaan 10, P.O. Box 283 (HQ1059), 1000 EA, Amsterdam, The Netherlands.
E-mail: nam.kyoo.boots@nl.abnamro.com
This work was done while N.K. Boots was with Departments of Econometrics, Vrije Universiteit, Amsterdam.
Correspondence to: M. Mandjes. E-mail: michel@cwi.nl

1. Introduction

With the advent of high-speed packet-based network technologies – such as the *Internet Protocol* – an accurate prediction of the achievable performance becomes extremely useful. In the end, the *Quality of Service* (mostly expressed in terms of performance metrics like packet loss, delay, throughput) is what the users experience, and in this sense it determines the success of the service. Consequently it is of major interest to get insight into the factors that affect the network performance. One could think of the influence of the characteristics of the traffic offered to the system (particularly its 'variability' in time, commonly called *burstiness*), as well as the features of the network and its network elements (buffer sizes, link speeds, routing).

In performance prediction through mathematical modeling, a crucial role is played by the *traffic model*. The common procedure is the following. First traffic measurements are done. These are used to develop a traffic model – such a model is usually phrased in terms of a stochastic process. Finally, it is calculated what performance is realized if this traffic stream feeds into the network – the type of models used are usually queueing models.

Long-range dependence. In other words, misspecifications of the traffic model might cause inaccurate performance predictions. This explains why the discovery of long-term dependences (in the beginning of the 1990's – a key paper is Leland *et al*. [23]) raised considerable concerns. Before, it was generally accepted that *short-range dependent* (SRD) source models captured all essential features of network traffic, i.e., models in which the correlation function of the arrival process decays exponentially in time. *Long-range dependent* traffic, however, would require a slowly (for instance polynomially) decaying correlation function.

After the discovery of LRD, one wondered what made network traffic behave like this. A key result here states that the aggregate of a large number of sources with heavy-tailed on- and/or off-times looks like an LRD process [41]. Then it can be argued that an individual user transferring files tends to resemble a heavy-tailed on/off fluid source, due to the heavy-tailed distribution of file sizes. The aggregate of many users leads to LRD traffic [9,41].

Queueing results. For SRD sources an extensive body of queueing results is available. Usually one considers the buffer overflow probability, but due to the constant service rate of the queue, this can be translated easily into

1434-8411/04/58/02-101 $ 30.00/0

the probability that the delay is longer than a specific threshold. Notably, for SRD input the overflow probability decays (roughly) exponentially. The case of statistically identical phase-type on/off fluid sources was explicitly solved [1,22]: the overflow probability can be expressed in terms of the solution of an eigensystem.

However, for queues fed by a superposition of LRD sources hardly any results were available. Therefore, during the past, say, five years, the focus shifted towards those models. Explicit solutions of the buffer occupancy distribution are not known; considerable attention was paid to *large-buffer asymptotics*. The earliest results [7,31] were on GI/G/1 queues with heavy-tailed service times. For the case of multiple on-off sources partial results were derived, see e.g. [4,5,20]. Remarkably, in these cases the large-buffer asymptotics inherit the heavy tail of the service times (for GI/G/1) or on-times (for superposition of on/off sources); the tail is in general even heavier than the service times or on-periods themselves.

Impact on performance. The above results suggest that LRD input (rather than SRD input) indeed leads to performance degradation. However, there are two fundamental objections against the use of large-buffer asymptotics. In the first place, the convergence is typically slow: only for extremely large buffers the asymptotic is accurate. In the second place, not for all applications the regime of large buffers is the most relevant one. Particularly for real-time applications smaller buffers (or: smaller delay thresholds) are more important.

This last point was well-taken in a couple of more practically oriented articles by Ryu and Elwalid [34], Heyman and Lakshman [18], and Grossglauser and Bolot [17]. Based on mathematical modeling and experiments with real traffic traces (the so-called Bellcore trace, VBR (Variable Bit Rate) MPEG video, etc.), they arrive at a common conclusion: as long as delay requirements are in some sense stringent, only the correlation structure on shorter time-scales plays a role. Long-term correlations do not have a significant impact, and hence SRD models can be used.

Analysis. In our study, we succeed in getting a theoretical handle on the result found in [17,18,34]. Our analysis consists of two steps: First we present a versatile queueing model of traffic multiplexed at a router, and then we synthesize a number of strong existing structural results. Then we use these results to assess the influence of the sources' characteristics on the performance for realisic scenario's. In greater detail, our contribution is the following.

1. Our versatile queueing model is the following. We have n i.i.d. sources, feeding into a FIFO queue with buffer B and link rate C. A generic source model is considered: traffic generated by an individual user is modeled as a fluid on/off pattern with generally distributed on- and off-times. Notice that this model covers both LRD (choose heavy-tailed on-times and/or off-times) and SRD input.
   A crucial point is that we do not focus on large-buffer asymptotics but rather on *many-sources asymptotics*: we let the aggregation level n grow large, and at the same time the resources are scaled accordingly: $B \equiv nb$ and $C \equiv nc$. Notice that in many practical applications the assumption of many sources is considerably more realistic than the assumption of large buffers (or equivalently: large delay tolerance).
   The asymptotics developed by Botvich and Duffield [3] state that the overflow probability decays exponentially in n. In this paper a major role is played by the resulting decay rate, particularly as a function of b. The results from [3] enable us to calculate the *loss curve* $I(\cdot)$, i.e., the decay rate as function of b.
   A disadvantage of the use of the Botvich-Duffield result is that it is implicit, in that the value of $I(b)$ is hidden behind a variational problem. Therefore we consider explicit characterizations of the loss curve for small and large b [3,25,26]. Here the fundamental difference between LRD and SRD input comes to the surface. Crucially, for small buffers an *insensitivity result* holds: $I(b)$ depends on the on- and off-times only through their means – in other words, SRD and LRD sources (with the same mean on- and off-times) behave roughly identically. For large buffers however, there is a distinction between SRD sources, where $I(b)$ is (at least) linear in b, and LRD sources, where $I(b)$ is sublinear (for instance like b^ρ, with $\rho \in (0,1)$, or like $\log b$). Hence, for LRD input the decay of the loss probability in the buffer size is slower than exponential; this is in stark contrast with SRD input, which has at least exponential decay. For the regime of large b we also have insightful properties of the behavior of the sources during the queue's path to overflow, which again indicate the fundamental differences between SRD and LRD input.
   Although our model is versatile (covering a broad range of traffic types), it of course has a number of less realistic properties. In practice, traffic that is multiplexed on a network will be heterogeneous (rather than homogeneous), and it will traverse a concatenation of links rather than just one. Also, in our model (unlike TCP), the rate the users send at does not adapt to the available resources in the network. We will detail these drawbacks, and argue why our model still captures the essential features.
2. Armed with the characteristics of the loss curve, we are in a position to assess the impact of LRD on the experienced performance. We select a number of scenarios of applications (voice, video, file transfer, web browsing, ...), and use traffic parameters that appeared in the literature and application-dependent delay requirements.
   Then for any value of the maximum delay (that may be exceeded by no more than a small fraction of the packets) we can compute how many sources of a specific type can be admitted on a link with given rate. We examine to what extent this number is affected by the shape of the on- and off-time distributions (keeping the mean on- and off-times fixed).

Our conclusions on the impact of LRD on the performance can be summarized as follows: (i) If the delay threshold is strict, there is hardly any difference between LRD and SRD input. For delay requirements in an 'intermediate' range, the probabilistic law of the input streams plays a role, but the 'heaviest tails' do not necessarily lead to the worst performance. Only for very tolerant delay requirements the large buffer results, as mentioned above, kick in. (ii) The level of aggregation is a significant factor. If the ratio between the link rate and the peak rate of a single source is high (and the sources are not too bursty), a high utilization can be achieved, while at the same time the delay requirements are met; this holds even if the delay requirements are stringent. Consequently, in traffic engineering one could use tight delay requirements, corresponding to the (insensitive) small buffer situation, while still running the system at a fairly efficient level.

This paper is organized as follows. Section 2 concentrates on the multiplexing queueing model, and describes the main properties of the loss curve. Section 3 applies this modeling to assess the impact of the model parameters on the realized performance. Section 4 reflects on important caveats regarding LRD traffic – particularly illegitimate reversal of limits leads to misleading results. Section 5 concludes.

2. The shape of the loss curve

This section presents the mathematical model underlying our analysis. The model and preliminaries are provided by Section 2.1. Structural results on the loss curve are given in Sections 2.2 and 2.3. The main contribution of this section is that we give a complete overview of the relevant results that provides (on an abstract level) important insights in the fundamental differences between LRD and SRD input. We do so by combining theoretical results that appeared in [3] and our previous work [25,26]. In Section 2.4 we comment on the influence of the correlation structure of the arrival process on the shape of the loss curve.

2.1 Model and preliminaries

Model. We consider traffic from n on-off fluid sources feeding into a buffered resource. This resource is modeled as a queue with constant depletion rate C. The traffic rate of each source alternates between on and off; during the on-times traffic is generated continuously at a (normalized) peak rate of 1. The activity periods constitute an i.i.d. sequence of random variables, each of them distributed as random variable A with values in $\mathbb{R}_+$. The silence periods are also an i.i.d. sequence, distributed as random variable S with values in $\mathbb{R}_+$. Both sequences are mutually independent. Define also

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

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 [14]. 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

$$F_{A^\star}(x) := \mathbb{P}(A^\star \le x) = \frac{1}{\mathbb{E}A}\int_0^x \mathbb{P}(A > y)\mathrm{d}y;$$

$S^\star$ is defined analogously.

Performance measures. We are interested in the probability of the buffer content exceeding level B, denoted by $p(B, C)$. Using the constant depletion rate, it is not hard to see how this performance metric can be translated into the probability that the queueing delay exceeds some prespecified threshold. Unfortunately, only in a few special cases $p(B, C)$ can be evaluated explicitly. This motivates why we resort to asymptotics.

In this work we choose the asymptotic regime in which the number of sources, say n, grows large. At the same time we rescale the resources : $C \equiv nc$ and $B \equiv nb$. This scaling was first introduced by Weiss [39] and has proven to be very powerful, see e.g. [3,8,36]. We believe that this scaling is quite natural: network elements of current packet networks are usually fed by many relatively small flows. In any case, the asymptotic regime of many sources seems to be more realistic than the regime of large buffers, as the latter regime is not appropriate for delay-sensitive applications.

We assume that the system is stable and non-trivial: $p < c < 1$. 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):

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

We call this curve the *loss curve*. The key result on $I(b)$ is given below in Theorem 2.3.

Theorem 2.3 describes the many sources asymptotics for general b. Botvich and Duffield [3] proved it under fairly general conditions, whereas related results were derived in [8,24,36]. The result that we use in this paper is

a slight variation of [24], that requires the following mild additional assumption.

Assumption 2.2 *Assume* $\inf_{t>0} I_t(b)$ *is a continuous function of* b, *where*

$$I_t(b) := \sup_\theta \left(\theta(b+ct) - \log \mathbb{E}e^{\theta A(t)}\right).$$

Theorem 2.3 [Loss curve for general b**]** *Under Assumptions 2.1 and 2.2,*

$$I(b) = \inf_{t>0} I_t(b) = \inf_{t>0} \sup_\theta \left(\theta(b+ct) - \log \mathbb{E}e^{\theta A(t)}\right). \quad (1)$$

For the proof of Theorem 2.3 we refer to Mandjes and Borst [26]. Informally, the following exponential approximation applies:

$$p_n(b,c) \approx e^{-nI(b)}, \quad n \text{ large}.$$

Discussion. Wischik [42] provides useful insight into the heuristics behind Theorem 2.3. He phrases $I(b)$ as an optimization over all paths of the buffer leading to overflow. His reasoning shows that the optimizing value of t, in the sequel denoted by $t_b^\star$, can be interpreted as the typical duration from the epoch the buffer starts to fill until overflow, given that this busy period leads to overflow. Therefore, we will call $t_b^\star$ the *most likely time to overflow*. The most substantial drawback of Theorem 2.3 is its intransparency: its value is concealed behind the inf sup program. This explains the interest in simple approximations of $I(b)$ for small and large b. In the next two subsections we review the approximations for small buffers (found in [26]) and large buffers (see [3,25]). For large buffers the loss curve is strongly affected by the distributions of the on- and off-times – we explain in detail the intuition behind this.

2.2 Small buffers: insensitivity in the shape of the distribution

The small buffer implies that the state of any individual source is not likely to change often during the trajectory to overflow, simply because the time to overflow is small. This is formalized in the next assumption, which is satisfied for a broad class of on- and off-time distributions – in fact it is enough that the corresponding densities are bounded.

Assumption 2.4 *The probability that the state (i.e., on or off) of any individual sources makes two or more transitions in an interval of length* t *is* $O(t^2)$, *where* $t \downarrow 0$.

Now define the following two constants:

$$\alpha(c) := c\log\left(\frac{c}{p}\right) + (1-c)\log\left(\frac{1-c}{1-p}\right),$$

and

$$\beta(c) := 2\sqrt{\left(\frac{c}{\mathbb{E}A} + \frac{1-c}{\mathbb{E}S}\right)\log\left(\frac{c}{1-c}\cdot\frac{\mathbb{E}S}{\mathbb{E}A}\right) - 2\left(\frac{c}{\mathbb{E}A} - \frac{1-c}{\mathbb{E}S}\right)}.$$

The following theorem is proved by Mandjes and Kim [26].

Theorem 2.5 [Loss curve for small b**]** *With* α *and* β *defined above, for small* b,

$$\lim_{n\to\infty} \frac{1}{n}\log p_n(b,c) = -\alpha(c) - \beta(c)\sqrt{b} + O(b), \quad (2)$$

under Assumptions 2.1, 2.2, and 2.4.

- *Influence of distributions*. Importantly, Theorem 2.5 states that the exponential decay rate depends on the distribution of the on and off-times, *only through their means* $\mathbb{E}A, \mathbb{E}S$. In other words, for given means, and small b, is *not* important whether the distributions have heavy tails or exponential tails. Consequently, the small buffer results found by Weiss [39] for exponential on/off sources generalize to general on/off sources.
- *Multiplexing gains*. For small buffers the loss curve $I(b)$ increases like $\sqrt{b}$. This means that for small buffers the overflow probability decreases very fast, so there is a large 'marginal benefit' of an additional unit of buffer space.
  It was already widely recognized that small buffers were useful to absorb traffic fluctuations at the packet level (that are due to the asynchronous arrival of packets). However, the shape of the loss curve for small b states that even if traffic is modeled as fluid (and hence the packet level is ignored), it is worthwhile to have a small buffer.
- *Path to overflow*. The time to overflow is proportional to $\sqrt{b}$; the proportionality constant is a straightforward function of $\mathbb{E}A, \mathbb{E}S$, and c, see Eq. 11 of [26]. As for the case with Exponential on- and off-times [39], the trajectory to overflow looks like a hyperbolic cosine.

2.3 Large buffers: linear and sublinear loss curve

For large buffers no insensitivity result applies. We recapitulate two results, namely the result for light-tailed on-times (giving rise to SRD input) by Botvich and Duffield [3] and the result for heavy-tailed on-times (giving rise to LRD input) by Mandjes and Borst [25]. These results state that the shape of the distribution of the activity period essentially determines the shape of the loss curve for large b.

We need a formal classification of probability distributions. Particularly, we rely on the concept of *subexponential* distributions, defined below in Definition 2.6 and

reviewed in detail in the appendices of [5]. The heavy-tailed distributions we use in this paper are in the class of subexponential distributions. We also define the class of subexponentially varying distributions.

Definition 2.6 [Heavy-tailed distributions] *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. With $F_X(\cdot) := \mathbb{P}(X \le x)$, we say that X has a subexponential *distribution, or $F_X(\cdot) \in \mathcal{S}$. Suppose the function $v_X(\cdot) := -\log \mathbb{P}(X>t)$ 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 $F_X(\cdot) \in \mathcal{V}$.*

Unfortunately, the exact relation between the classes $\mathcal{S}$ and $\mathcal{V}$ is not clear. The most important heavy-tailed distributions (like Pareto, Lognormal, and Weibull) are in both of them. A well-known implication [5, Lemma 7.2 and 7.3] of $F_X(\cdot) \in \mathcal{S}$ is that for all positive ϵ,

$$e^{\epsilon t}\mathbb{P}(X>t) \to \infty, \quad \text{as } t\to\infty. \tag{3}$$

– *Light tails*. First we review the case of light-tailed on-times, due to [3]. Define

$$\theta^\star := \sup\left\{\theta : \lim_{t\to\infty} t^{-1}\log \mathbb{E}e^{\theta(A(t)-ct)} \le 0\right\}. \tag{4}$$

Theorem 2.7 [Loss curve for large b – light-tailed on-times] *With $\theta^\star$* defined in (4),

$$\lim_{b\to\infty} I(b) - \theta^\star b = \nu,$$

under Hypotheses 1.(i)–(iv) of [3], *and assuming that $\nu := -\lim_{t\to\infty}\log \mathbb{E}e^{\theta^\star(A(t)-ct)}$ exists.*

The crucial assumption here is Hypothesis 1. (iii) of [3], i.e., there exists a positive θ such that $\mathbb{E}e^{\theta(A(t)-ct)} < 1$ for all t large enough. For on-off sources with subexponential on-time, because of (3), for $\theta>0$ and t large:

$$\mathbb{E}e^{\theta A(t)} \ge p\,\mathbb{P}(A^\star>t)\,e^{\theta t} \ge e^{\theta ct};$$

here we focused on the probability that the source is on at time 0 and stays on during $[0,t]$. Therefore Theorem 2.7 is not applicable if the bursts are heavy-tailed.

– *Heavy tails*. The following theorem, from [25], covers the case of heavy-tailed on-times.

Theorem 2.8 [Loss curve for large b – heavy-tailed on-times] *If $F_{A^\star}(\cdot) \in \mathcal{S}\cap\mathcal{V}$:*

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

with $v(t) := -\log\mathbb{P}(A^\star>t)$, under Assumptions 2.1 and 2.2.

- *Influence of distributions*. If the on-times are not heavy-tailed, Theorem 2.7 applies, even if the off-times are heavy-tailed. Hence $I(b)$ is asymptotically linear in b, implying that the decay of the loss probability in the buffer size is essentially exponential. An alternative expression for $\theta^\star$ is

$$\sup\left\{\theta : \mathbb{E}e^{\theta A\ (1-c)}\mathbb{E}e^{-\theta Sc} \le 1\right\},$$

cf. [13,40]. In other words, in this regime, the loss curve depends on the entire distributions of A and S. If the on-times are heavy-tailed, according to Theorem 2.8, $I(b)$ more or less looks like $v(b)$. Hence $I(b)$ is sublinear, i.e., the decay of the overflow probability is slower than exponential. More precisely, $I(b)$ looks like $\log b$ for Pareto on-times, and like b^ρ, $\rho\in(0,1)$, for Weibull on-times. Notice that the asymptotics depend on the distribution of S only through its mean.
- *Multiplexing gains*. It was already mentioned that for small b the overflow probability decreases fast in b. From Theorems 2.7 and 2.8, we conclude that this marginal benefit is smaller for larger b, since the loss curve increases only linearly or even sublinearly.
Notice that Theorem 2.7 is a more accurate than Theorem 2.8, in that the former gives a function $f(\cdot)$ such that $I(b)-f(b)$ tends to a constant, whereas the latter gives a function $g(\cdot)$ such that $I(b)/g(b)$ tends to a constant. Theorem 2.7 nicely describes (for light-tailed input) the multiplexing gain that can be achieved on top of bufferless multiplexing, as opposed to the crude 'effective bandwidth approximation' $I(b)\approx\theta^\star b$, see the introduction of Botvich and Duffield [3] and Choudhury, Lucantoni, and Whitt [6].
- *Path to overflow*. The theoretical results of this section enable us to get a better qualitative understanding of the most likely way for buffer overflow to occur.
  – For systems with light-tailed on-times, detailed analyses are available. It is well understood that the sources must behave according to a different statistical law in order to fill a large buffer: The on-periods are longer and the off-periods shorter than during average behavior. More precisely: the on-

and off-times are *exponentially twisted*. Essentially, during the path to overflow, all sources behave according to the same 'new' statistical law, cf. the seminal paper of Weiss [39], and more recent articles [2,28].

– For sources with subexponential on-periods, the intuition behind the trajectory to overflow is completely different. During a path to overflow, there are two types of sources. A first group sends at peak rate the entire time from an empty system to overflow. Another group alternates between on and off, in such a way that they effectively contributing at mean rate (i.e., these sources behave according to their normal statistical law). Note that this in stark contrast with the behavior exhibited by light-tailed sources; as described above in that case all sources essentially behave in the same way, and alternate between on and off, effectively sending at a higher rate than their mean rate.
The validity of this intuition is supported by the following heuristic calculation. Consider the situation of n homogeneous on-off sources with $F_{A^\star}(\cdot) \in \mathcal{S} \cap \mathcal{V}$. Let us follow the above intuition, and let K be the number of sources that send at peak rate. An approximation for the loss probability is

$$p(B,C) \approx \max_{K:K+(n-K)p>C} \mathbb{P}\left(A^\star > \frac{B}{K+(n-K)p-C}\right)^K,$$
$$K \in \{0,\ldots,n\}.$$

Putting $K \equiv nk$,

$$\frac{1}{n}\log p_n(b,c) \approx -\min_{k:k+(1-k)p>c} k \cdot v\left(\frac{b}{k+(1-k)p-c}\right) \approx -\min_{k:k+(1-k)p>c} k\cdot(k+(1-k)p-c)^{-h} v(b). \tag{5}$$

The minimum is reached for

$$k^\star = \min\left\{\left(\frac{c-p}{1-p}\right)\left(\frac{1}{1-h}\right), 1\right\}. \tag{6}$$

Inserting $k^\star$ into (5) this indeed directly leads to the decay rate given in Theorem 2.8. Notice that $k^\star$ can be interpreted as the fraction of sources that send at peak rate during the entire path to overflow.

– *Examples*. We give three examples of sources with essentially different trajectories to overflow. The off-times are assumed to be exponentially distributed.

1. Light-tailed on-times. A and S are exponentially twisted. Following [2], for large b,

$$t_b^\star \approx \frac{\mathbb{E}\bar{A}+\mathbb{E}\bar{S}}{(1-c)\mathbb{E}\bar{A}-c\mathbb{E}\bar{S}},$$

where $\mathbb{E}\bar{A} := \dfrac{\mathbb{E}A\, e^{\theta^\star A(1-c)}}{\mathbb{E}e^{\theta^\star A(1-c)}}$

and $\mathbb{E}\bar{S} := \dfrac{\mathbb{E}Se^{-\theta^\star Sc}}{\mathbb{E}e^{-\theta^\star Sc}}$.

2. Pareto on-times. 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 an $\alpha > 1$. As $h=0$, according to (6) a fraction $k^\star = (c-p)(1-p)^{-1}$ send at peak rate essentially during the entire path to overflow, whereas the remaining fraction $(1-c)(1-p)^{-1}$ contribute at mean rate p (by alternating between on and off with their 'normal' statistical law). An easy calculation gives aggregate input rate c. In other words: if $h=0$, then the net input rate will be only slightly larger than 0. This suggests that [25] $t_b^\star$ should grow faster than linearly in b. In fact, Mandjes and Borst [25] show that $t_b^\star = bf(b)$, with $f(\cdot)$ such that $\log(bf(b))/f(b) \to 1$ (with b large). Thus $f(b)$ is clearly smaller than polynomial, but larger than a constant. It is easily checked that for A Lognormal we have similar behavior.
3. Weibull on-times. Here A has a cumulative distribution function $\exp[-t^\beta]$, which leads to a $v(\cdot)$ function which is regularly varying of index β, with $\beta \in (0,1)$. From (6) it is seen that the net input rate is positive, thus leading to a time to overflow that is essentially linear in the buffer size, with $t_b^\star k^\star(1-c) \approx b$. If h is close to 1, then all sources will have long bursts (as $k^\star = 1$).

To illustrate the influence of the distributions, we conclude this section with characteristic graphs of $I(b)$ and $t_b^\star$ as functions of b. In Figure 1 we compare light-tailed (Geometric), Pareto, and Weibull on-times. For numerical ease, we use slotted time; consequently the $I(b)$ curve does not quite look like a square root for small b. It can be verified that for large b, $I(b)$ is indeed linear for Geometric bursts, log-like for Pareto, and polynomial for Weibull. Notice the superlinear behavior of $t_b^\star$ for Pareto on-times.

2.4 Correlations and concavity

A general conjecture is that, if the packet arrivals are negatively (positively) correlated on time scale $t_b^\star$, then the loss curve is convex (concave) at b. Empirical motivation for this conjecture can be found in [3, Section 4.4]. There a discrete-time queue is considered, fed by

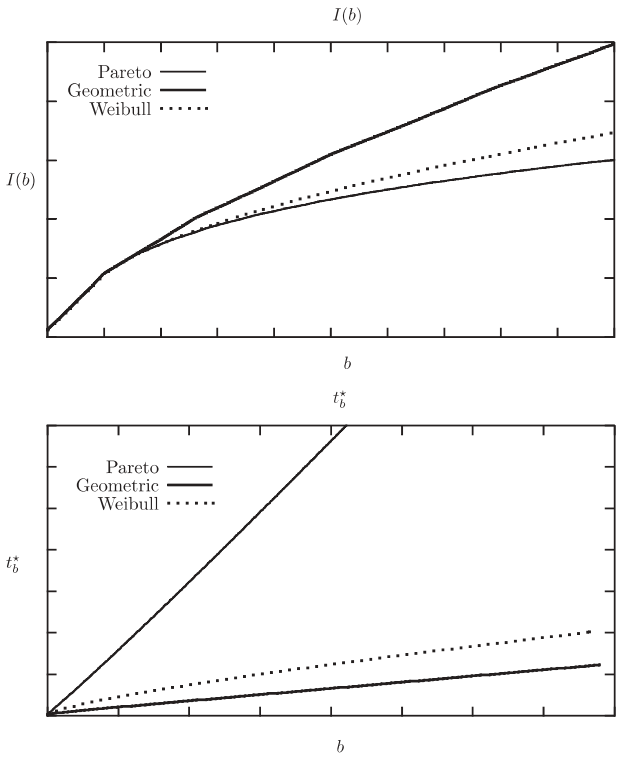


Fig. 1. $I(b)$ and $t_b^\star$, as functions of b.

sources with Geometric(q_1) on-times and Geometric(q_2) off-times. They found that depending on the correlation structure, the loss curve has a convex or concave shape. More precisely, for $q_1 + q_2 > 1$ (negative correlation) they showed convexity, for $q_1 + q_2 < 1$ concavity (positive correlation).

In the cases described in the previous subsections the loss curve is concave, due to the positive correlations of the inputs that satisfy the assumptions of Theorems 2.7 and 2.8. However, it is possible to construct on-off fluid sources with negative correlations, for instance by taking deterministic on and off-times. In the literature significant attention has been paid to this type of 'adversarial traffic' [15].

From the formulas reviewed in the previous subsections we also conclude that the level of correlation determines the level of concavity. For $b = 0$ the curve is highly concave (second derivative is ∞), for larger b (and consequently longer associated time scale) the concavity is less pronounced. In the light-tailed case the concavity vanishes: the loss curve has a linear asymptote. This is in line with the observation that for light-tailed activity periods there is indeed hardly any correlation left on the relevant time-scale (which is proportional to b), due to the short-range dependent character of the sources. In the heavy-tailed case the loss curve could be still quite concave ($\log b$ for Pareto, b^β for Weibull), because on the relevant time scale still considerable positive correlations exist.

3. Numerical evaluations

Section 2 indicated that for small buffers LRD hardly affects queueing performance, whereas for large buffers it does. Hence, it is of crucial importance to identify which of these two regimes applies in realistic situations. To that end, our approach is the following. We first list a number of relevant applications (voice, video, file transfer, web browsing, etc.). The corresponding traffic characteristics (in terms of our on-off model) and performance requirements are identified from empirical studies, e.g., [9,32]. Then we compute, for different values of the link rate C, how many flows can be accepted without violating the performance criterion, varying the shape of the distributions (but leaving the mean on- and off-times constant). Clearly, this statistic gives important insight into the impact of the traffic characteristics.

This section is organized as follows. We start by presenting the related literature in Section 3.1, and indicate where we depart from their approach. Then we describe in Section 3.2 the traffic scenarios and performance requirements. In Section 3.3 we assess the impact of the traffic characteristics for the described traffic scenarios. Section 3.4 presents the conclusions.

3.1 Literature on the impact of LRD

Before we present our own approach, and its results, we first briefly review a number of important contributions on the impact of LRD.

Ryu and Elwalid [34]. In this paper attention is focused on multiplexed real-time VBR video sources (with purposes like video conferencing, etc.). The main conclusion is that the long-term correlations do *not* affect the performance – short-term correlations are dominant, and therefore Markov modeling is adequate. The main reason behind this lies in the rather strict delay requirement (in the order of 20–30 ms per queue) imposed by real-time video. The performance metric used is the probability that the delay exceeds the upper bound mentioned above, which can be translated into the probability that the buffer content exceeds some specific level. It can be argued that the strict delay constraint implies, loosely speaking, that the buffer has little memory, such that long tails cannot have a significant impact. The analysis relies on the notion of critical time-scale, i.e., the number of time-lags that contribute to the buffer overflow probability – the authors show that even in the presence of LRD this number is small.

Heyman and Lakshman [18]. The authors also consider real-time VBR video, and not surprisingly, the conclusion of the paper is similar to the one of [34]. LRD does not affect the buffer occupancy distribution significantly, and Markovian models suffice to accurately predict performance. Only knowledge of the mean and variance of the marginal distribution and the lag-1 autocorrelations are required. The authors advocate the use of the (short-range dependent model) DAR(1), i.e., a discrete autore-

gressive model of order 1. The authors present additional results in [19].

Grossglauser and Bolot [17]. The authors conclude that the amount of correlation required is determined by the time-scales that are typical for the system under consideration. This justifies the use of Markov models (or self-similar models, as long as they have the right correlation structure up to the 'correlation horizon'). This in line with the findings from [18,34]. Grossglauser and Bolot [17] consider the packet loss rate rather than the probability that the delay exceeds some critical value. The authors also conclude that, in order to decrease the loss rate, it is much more efficient to adjust the marginal distribution of the rate than to use large buffers. The authors propose a modulated fluid traffic model of which a special case is constituted by a superposition of on-off sources.

Evaluation. The studies mentioned [17,18,34] have a strongly empirical character, supported by mathematical modeling. Both [18] and [34] exclusively address real-time video, although the same question ('what is the impact of LRD?') is of great importance for other applications. Below we will define a broader set of applications.

The interesting point of [34] is the notion of critical time scale, i.e., the number of time-lags that contribute to the overflow probability. The authors use large deviations theory to support this – the role of the critical time scale is comparable to $t_b^\star$ in our analysis, and the correlation horizon identified in [17].

As noted above, the model of [17] covers on-off sources. However, they assume that the distributions of the bursts and the silences are *identical*, which seems to be quite restrictive. Clearly, our model does not have this constraint.

The performance metric used in [17] is the packet loss ratio, instead of the probability of exceeding some delay level. This does not seem to be so adequate, as the authors mention that the buffer that they consider is so large that packets can have a delay of a few seconds. Evidently, in applications like real-time applications, it is usually not desirable that packets experience delays of that order. Therefore we prefer the probability of the delay exceeding some predefined bound (considerably smaller than a few seconds).

From the above, we conclude that there is a need for a unified modeling that covers a broader set of applications (apart from video also applications like audio, file transfer, etc.). In the next subsection we detail the approach that is followed in the present paper.

3.2 Approach

We will use the results of the previous section to shed some light on the impact of LRD. For the sake of convenience we choose slotted time. An important advantage of discrete time is that it is easy to evaluate the moment generating functions recursively (as described in Appendix A). At the same time, Theorem 2.3 goes through. Also the theorems on small and large buffers are essentially still valid, given that the number of packets per burst is large (because then there is little difference between the discrete-time model and the fluid model).

Performance measure. The metric we use is the probability p_D of the packet delay exceeding some maximum – this probability must be small, typically in the order of 10^{-4}–10^{-5}. The delay probability can be translated into the overflow probability of Section 2: with delay requirement D, we must have that $p_n(cD, c) \leq p_D$.

For any value of the delay requirement D, we calculate the number of sources that can be admitted. If j is the number of sources that can be admitted to achieve this performance target, it is clear that j is an increasing function of D. As follows easily from Theorem 2.3,

$$j = \inf_{k\in\mathbb{N}} \sup_{\theta} \left(\frac{\theta(CD + Ck) + \log p_D}{\log \mathbb{E}\exp(\theta A(k))} \right). \tag{7}$$

Alternatively, we can choose n large (and as before $B \equiv nb$ and $C \equiv nc$), and δ such that $\exp[-n\delta] = p_D$. Then, $j = nJ(D)$, with

$$J(D) := \inf_{k\in\mathbb{N}} \sup_{\theta} \left(\frac{\theta(cD + ck) - \delta}{\log \mathbb{E}\exp(\theta A(k))} \right).$$

Note that in this case j rather than n denotes the number of sources. We call this curve (as a function of D) the *acceptance curve*. We will assess the impact of the distributions of the on- and off-times on the basis of this acceptance curve. In Appendix B we derive that it is, for small D, insensitive to higher moments of the activities and silences (just like the loss curve is).

Applications. The source models we present below do not intend to describe the stochastic behavior of the traffic flow as accurately as possible. However, we believe that the capture the essential features, such that we can draw general conclusions on the impact of the source characteristics. Table 1 summarizes the source models and performance requirements.

- *Scenario 1: Voice, non-real-time audio*. Due to its interactive character, voice has very strict delay requirements, typically in the order of a few ms per hop (router). We will consider voice with silence suppression, leading to on-off streams, with mean on and off-times in the order of a second. We will choose the parameters given in Sriram and Whitt [37]: activities of mean length 352 ms, and silences of mean length 650 ms, and a peak rate of 32 kbit/s. In the experiments below, we will vary the distribution of the activities and silences.
  As opposed to voice, non-real-time audio (for instance broadcast) does not impose severe delay constraints. One could think of a delay requirement up to 1 s per router. For reasons of simplicity we use the same traffic characteristics as those described above for coded voice.
- *Scenario 2: VBR video*. Several studies describe the statistical behavior of variable bit rate MPEG video – an overview of available models is given in Section 3.2 of Rose [33]. Jelenković, Lazar, and Semret [21] examine the traces from [33]. We will use a simplified

Table 1. Traffic source parameters.

Application	$\mathbb{E}A$ in s	$\mathbb{E}S$ in s	Peak rate (minimum rate) in kbit/s
voice/audio	0.352	0.650	32
VBR video	9.0	9.0	800 (400)
web browsing (i/ii/iii)	0.01/0.10/0.50	10	3000/300/60

version of the model of [21]: sources with two levels of activity, so-called scenes. These scenes have mean lengths of about 9 seconds, i.e., 18.75 so-called *Groups of Pictures* (GOPs), where a GOP corresponds to 0.5 s. The distributions of the scenes are i.i.d.: for both activity levels the density of the duration is Pareto with tail-parameter in the order of 2.5 (i.e., the probability of a scene exceeding level x roughly looks like $x^{-1.5}$). The traffic rate at the high activity level could be about 800 kbit/s (about $4 \cdot 10^5$ bits per GOP), and 400 kbit/s (about $2 \cdot 10^5$ bits per GOP).

Notice that the model presented in [21] is more accurate, as it identifies a fluid model as described above (on a somewhat longer time-scale), but also a detailed model for the short time-scale. Also they distinguish more than just two activity levels (four levels, with traffic rates of 230, 440, 680 and 1180 kbit/s). We believe however that our two-level model captures the main effects – notice that a queue fed by sources with two activity levels can be analyzed by the on-off models of this paper (by adjusting the peak rate and the link rate).

The delay requirements are quite stringent in case of real-time video (e.g., video conferencing), in the order of a few ms per hop; for broadcast video one could think of delays up to 1 s per hop.

- *Scenario 3: Web browsing*. A first important observation is that we should distinguish between packet delay and file transfer delay. The former is the delay an individual packet experiences, whereas the latter is roughly the delay between the request and the arrival of the last bit – we focus on the former. Packet delay requirements (per router) could be thought of as in the order of a few tens of ms.

  An important contribution to the distribution of file sizes is by Crovella and Bestavros [9]. They do extensive statistical analysis of WWW document sizes. They find a heavy-tailed distribution, where the tail of the complementary distribution function is of Pareto-type of index 1.0 up to 1.3, and with mean in the order of a few thousand bytes, see [9, Table 1]; comparable figures are given by Paxson and Floyd [32] for ftp connections on the Internet.

  As motivated in [9, Section 5] the traffic generated by a web browsing user could be described as an on-off source: the on-times are the transfers, the off-times are the think-times. The mean on-times depend of course on the peak rate at which the file arrives at the router. In our calculations below we will use (i) a high peak rate (in the order of 3000 kbit/s), (ii) a medium peak rate (in the order of 300 kbit/s), and (iii) a low peak rate (in the order of 60 kbit/s). We take a think-time with mean 10 s.

The idea is to plot the acceptance curve for the traffic profiles described above, for different distributions of the on- and off-times. We will do that for different values of the link rate C. Then, by visual inspection, for any delay requirement, we can conclude whether or not the specific distributions play a role or not. Based on Theorem 2.5 and Appendix B we expect that for small D there will hardly be any difference, but Theorems 2.7 and 2.8 indicate that for larger D the curves will diverge. The interesting question is: is there any difference at the practically relevant values of D?

Evaluation of the approach. The approach described above offers a unified framework for evaluation of the impact of LRD under realistic circumstances. Obviously the model does not capture all effects that play a role in practice. Below we present a number of these drawbacks, and argue why we think that our results are still applicable.

- *Traffic is homogeneous*. In practice, network traffic is composed from a number of heterogeneous sources. As explained in [3, p. 300] it is possible to calculate the loss curves of heterogeneous superpositions of sources. For the sake of clarity we will restrict ourselves in the numerical experiments below to homogeneous input. We expect that heterogeneous input will lead to similar figures.
- *No (feedback) rate control taken into account*. Real-time video and interactive voice are not likely to be transported by a feedback-based protocol. However for the other (non-real-time) applications there will be a role played by TCP – this protocol provides the sender with information on the state of congestion, on the basis of which he can adapt the rate at which he sends.

  In other words, for instance for a file transfer the pattern of packet arrivals is not on-off, on a detailed time-scale. However, as justified in [9, Section 5.2.1] on a somewhat longer time-scale the rough approximation by an on-off fluid source applies. Also, notice that it is possible to explicitly model TCP-like feedback control mechanisms by fluid models, e.g. the one examined by Mandjes, Mitra, and Scheinhardt [27]. For reasons of conciseness with did not use these models here.
- *Single link instead of network model*. We consider just a single link, but of course the *end-to-end* delay is the metric the user is interested in. One could take this into account by approximating the end-to-end delay by the sum of individual delays. Notice that this is quite con-

servative: if in all N queues the probability of a delay larger than D is ϵ, then (under an independence assumption), the probability that the end-to-end delay is larger than ND is much smaller than ϵ. In Van der Wal *et al*. [38] a method is explained how to generate more realistic estimates on the end-to-end delay.

- *Drawbacks of the traffic model*. Although the traffic model is rather generic (since the on and off-times are general), some remarks can be made here. As indicated above, for VBR video a model with more than just two levels could be more suitable. It should be emphasized that these more complicated models can in principle be treated by the same large deviations machinery.
  We also assume that sources are *stationary*: their statistical behavior is constant in time. Therefore, we could not deal with the cases described in [9], where Pareto-distributed file size are described with parameter 0.9.

3.3 Numerical results

In this section we present graphs of the acceptance curve $j(D)$ corresponding to the scenarios described in Section 3.1. This is done for different on-time distributions (as we saw in Section 2 that the distribution of the off-time does not really affect the shape of the loss curve). We also varied the link rate C, thus allowing different levels of multiplexing.

The on- and off-times are $\mathbb{N}$-valued random variables. Like in Figure 1, we choose the following distributions:

- Weibull(κ, τ) distribution ('moderately' heavy tail) with

$$\mathbb{P}(A=k) = e^{-[\tau(k-1)]^\kappa} - e^{-[\tau k]^\kappa} \quad (0<\kappa<1,\ \tau>0).$$

  In the experiments the κ, which determines the heaviness of the tail of the distribution, is fixed at 0.4. The τ is chosen such that the mean has the right value.
- Pareto distribution ('very' heavy tail) with

$$\mathbb{P}(A=k) = [\beta/(\beta+k-1)]^\alpha - [\beta/(\beta+k)]^\alpha$$
$$(\alpha, \beta > 0).$$

  The index parameter α determines the heaviness of the tail of the distribution. In the numerical examinations, α has a application-specific value. The β is chosen such that the distribution has the desired mean.
- Geometric(q_1) distribution (light tail) with

$$P(A=k) = (1-q_1)^{k-1}q_1 \quad (0<q_1<1).$$

  The mean of this distribution is $1/q_1$.

We take Geometric(q_2) off-times with

$$\mathbb{P}(B=k) = (1-q_2)^{k-1}q_2 \quad (0<q_2<1).$$

We evaluate three sizes of the link rate, namely 45 Mbit/s (aggregation level), and 150 Mbit/s and 600 Mbits/s (backbone). We take the delay exceedance probability p_D equal to 10^{-5} and we choose the packet size equal to 300 bytes. For our computations we use the convention that one Kbyte equals 1024 bytes.

As mentioned in Section 3.2, in our VBR video model traffic generated by a single source is not on-off, but rather is the rate alternating between two positive levels. For implementation purposes we normalize the peak rate to 1 (in the VBR video model we normalize the difference between the peak rate and the minimum rate to 1) and we scale the link rate, the minimum rate (in case of the VBR video model) and the mean of the on- and off-times accordingly.

Obviously, if there is no delay constraint the number of accepted sources is C/p, recalling that p is the mean rate. For this reason we also plot the line C/p in the pictures. As said before, a scheme for the computation of $\mathbb{E}\exp(\theta A(k))$ and the acceptance curve are given in Appendix A.

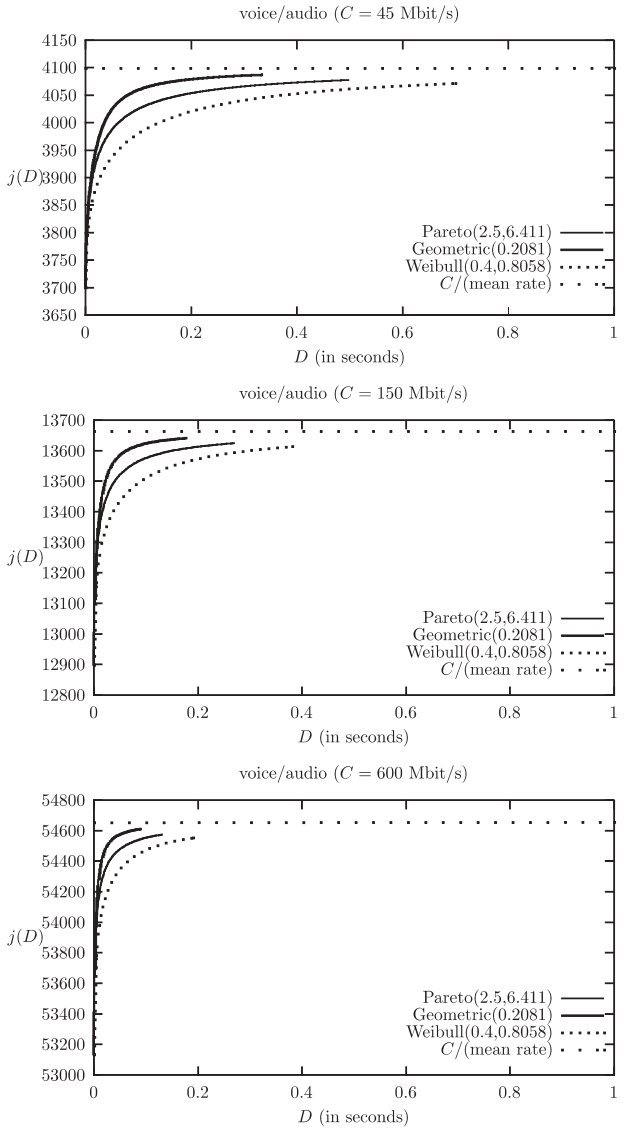
3.4 Discussion

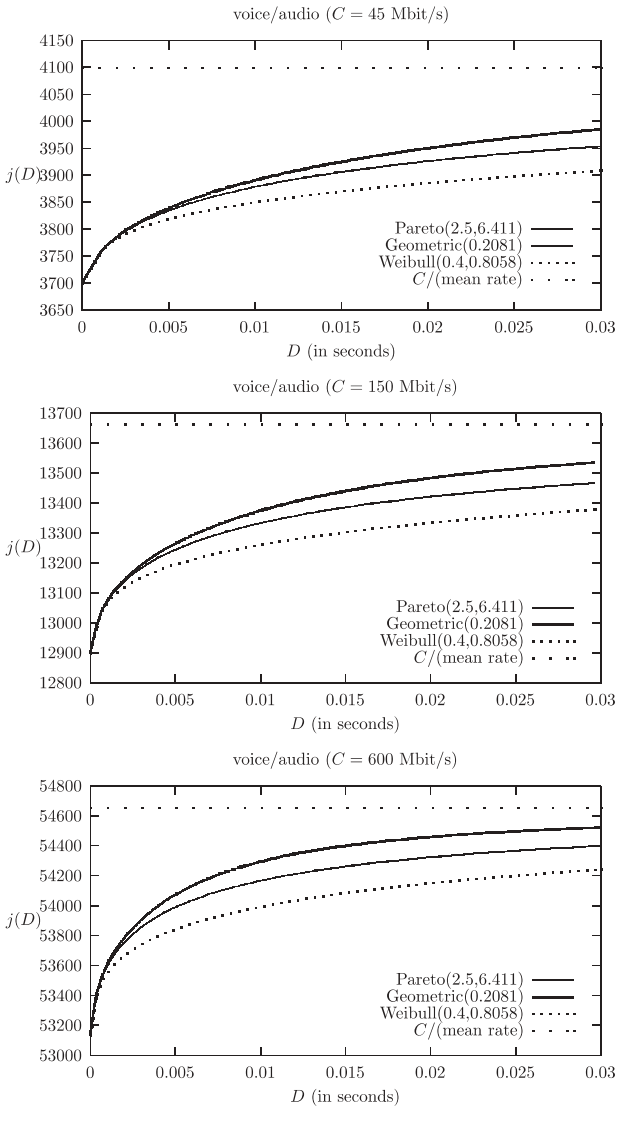
In this section we discuss the influence of the shape of the on-time distribution on the acceptance curve, as follows from the graphs in Section 3.3. We also conclude that the level of aggregation (i.e., the size of the link rate) is an important factor; below we comment on its impact.

The on-time distribution. A general conclusion is that the relevance of the on-time distribution strongly depends on the delay requirement D. As can be seen from the graphs, for stringent delay requirements the on-time distribution does not play a role at all; in fact we are in the small buffer regime. When the delay threshold increases the shape of the on-time distribution becomes more important. However, from the results for voice and video we conclude that here the heaviness of the tail is certainly not the only determining factor, since the graph of the Pareto distribution lies above the graph of the Weibull distribution (although for large enough delay this is no longer the case, according to Theorem 2.8). Apparently, detailed information on the shape of the distribution (not necessarily the tail) has significant impact.

For web browsing, presumably due to the large off-times (and consequently the large peak-to-mean ratio), the large buffer regime is reached rather quickly. Consequently the positioning of the graphs for the different distributions is as expected: Pareto is worse than Weibull, which is worse than Geometric.

Level of aggregation. From the above figures we can draw the following general conclusion regarding the level of aggregation. If the ratio between the link rate and the peak rate of a single source is high (and the sources are not too bursty), a high utilization can be achieved, while at the same time the delay requirements are met; this holds even if the delay requirements are stringent.



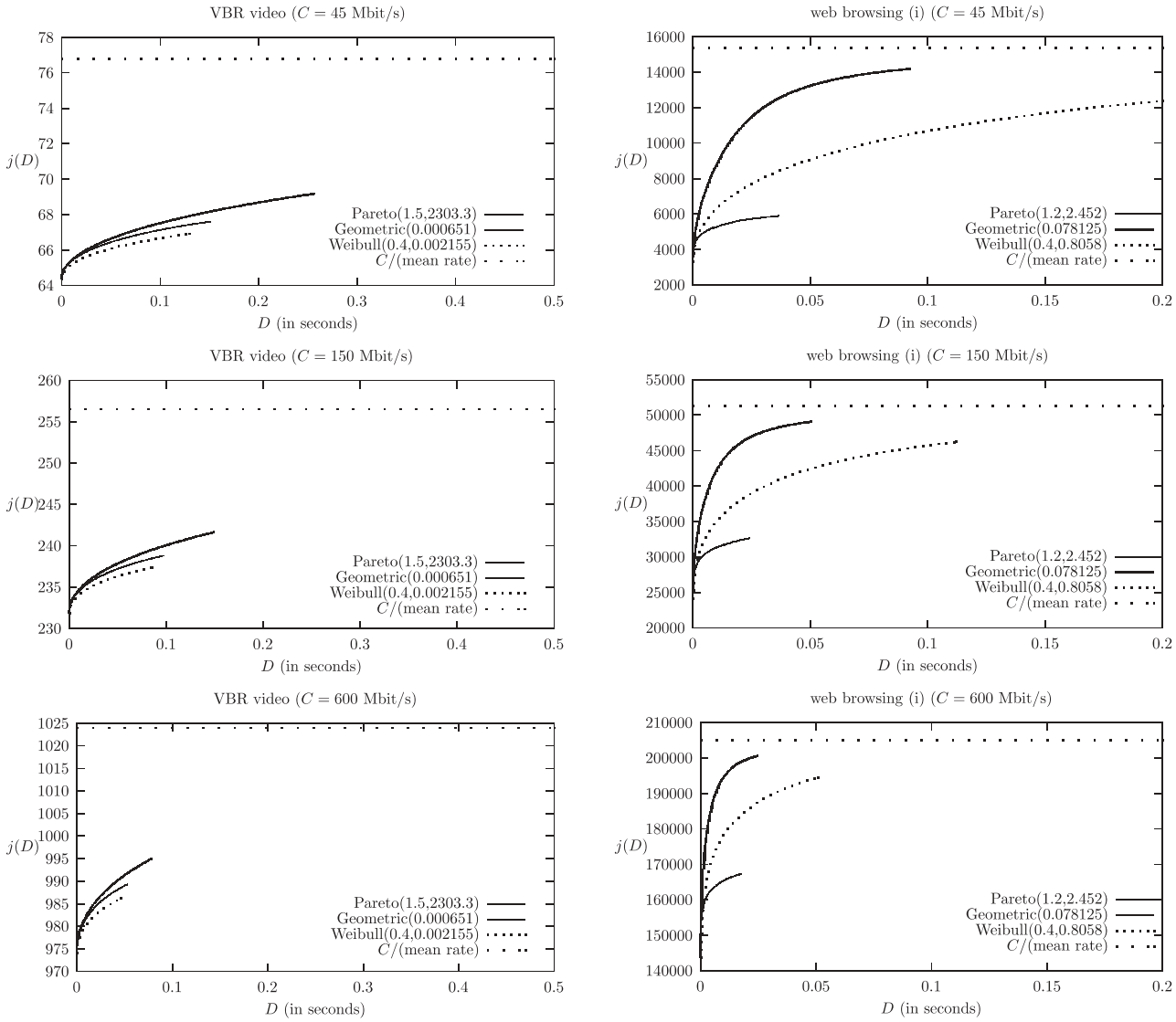


Consequently, in traffic engineering one could use tight delay requirements, corresponding to the (insensitive) small buffer situation, while still running the system at a fairly efficient level. One could even resort to the zero buffer case ('rate-envelope multiplexing') if the resulting efficiency is sufficiently high. From the graphs we conclude that the rate-envelope multiplexing utilization for voice and video is in the range 80–90%. Similar results hold for (ii) and (iii) of the web browsing model. Scenario (i) however leads to a poor utilization, particularly when $C = 45$ Mbit/s; this is due to the extremely high peak-to-mean ratio).

Clearly, if one is satisfied with the rate-envelope multiplexing utilization, distributions do not play a role at all. Only in case of a low level of aggregation (low link rates, for instance in the access network), in conjunction with (extremely) bursty input, this leads to a low efficiency. Then it could be worthwhile to exploit the buffer (equivalently: to allow for significant delay) in the traffic engineering guidelines. Unfortunately, this requires information about the on-time distribution, which is more detailed than just the mean. The graphs suggest that the efficiency can be increased considerably, even by a conservative choice of the on-time distribution (Weibull for voice and video, Pareto for web browsing).

4. On the impact of long-range dependence on network performance

With the theoretical results of Section 2, as well as the numerical results of Section 3 in mind, we are in a position to give a well-founded assessment of the influence of long-range dependence on network performance.



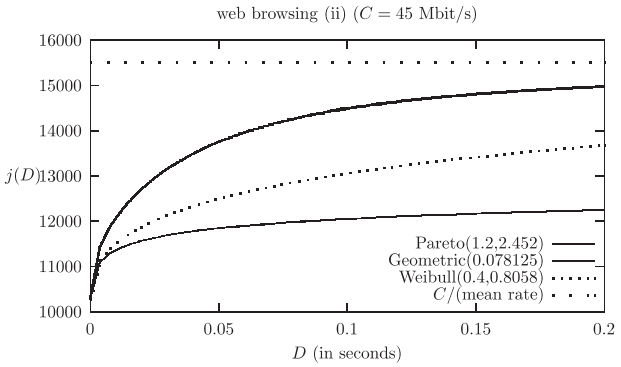
The structure we use in this section is the following. We phrase a number of statements that have some truth, but whose validity is more subtle. We detail the extent to which the statement holds, and where more care needs to be taken. Some of the arguments are perhaps already known in the literature; the text below is intended to give a complete account on this issue.

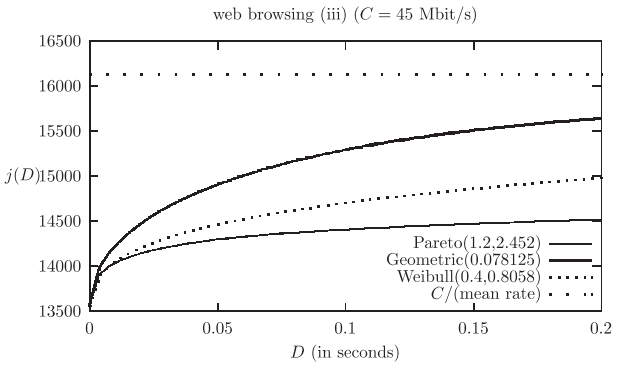
Claim 4.1 *If sources with heavy-tailed inactivity periods are multiplexed, this leads to performance degradation, in the sense that the tail of the queue length distribution is heavier than exponential.*

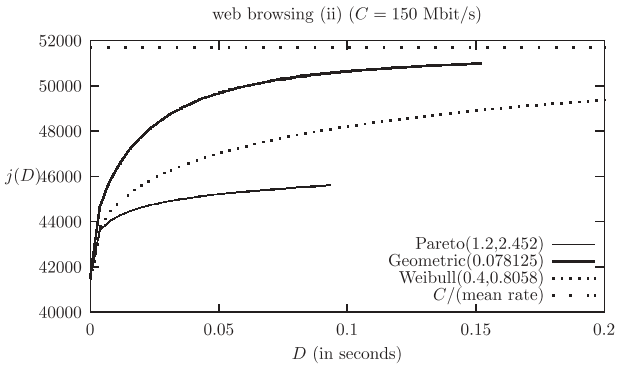
In the literature attention is paid to the tails of the distribution of *inactivity periods*. In many cases it was found that these are non-exponential – for instance Feldmann [16] describes that interarrival times of TCP connections can be accurately modeled by a heavy-tailed Weibull distribution.

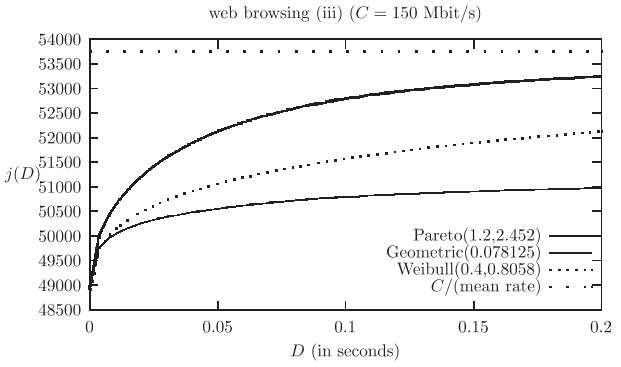
Now consider the situation that a large number of sources with heavy-tailed off-times are multiplexed. From the formulae of Section 2, it is not hard to see that this hardly affects the queue's tail behavior: (1) In case the on-times have a light tail, we get from Theorem 2.7 that the queue size distribution decays exponentially in the buffer size; (2) If on the other hand the on-times have a subexponential tail, Theorem 2.8 indicates that the queue size distribution mimics the heavy tail of the residual activity period; the off-time is represented just by its mean. We conclude that a possibly heavy tail of the off-time does not contribute to non-exponential tail behavior of the queue.

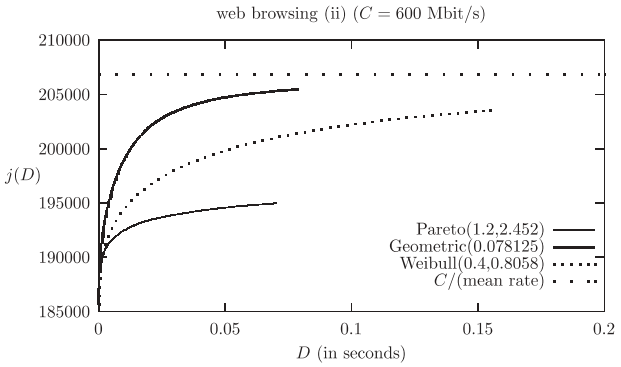
Claim 4.2 *The* Hurst parameter *is a valuable measure of long-range dependence. The higher it is, the fatter the tail of the queue size distribution, i.e., the worse the experienced QoS.*

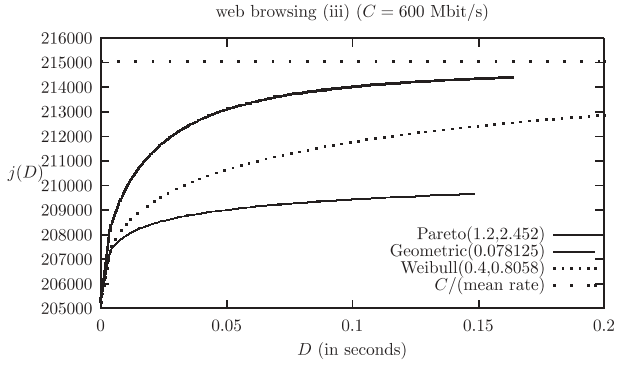












The statement is formally true: Consider fractional Brownian motion (FBM) $B_H(t)$ with Hurst parameter H, i.e., the Gaussian process with zero mean, stationary increments and correlation structure

$$\mathbb{E}\left(B_H(s)\cdot B_H(t)\right) = \frac{1}{2}\left(s^{2H}+t^{2H}-|s-t|^{2H}\right).$$

For a queue fed by this process it is known that the queue-length distribution has a Weibull-like tail with tail parameter $2(1-H)$. In other words, roughly the asymptotic relation

$$\mathbb{P}\left(\sup_{t>0} B_H(t) - Ct > B\right) \approx \exp\left(-\kappa B^{2(1-H)}\right)$$

applies [29,30]. In other words, indeed, a higher H leads to performance degradation.

However, a number of limit results that appeared in the literature might lead to some confusion here. Consider on-off sources of which either the on-times are of Pareto-type (of index α_{on}) or the off-times of Pareto-type (of index α_{off}), or both. Loosely speaking, in [41] it was show that the aggregation of many of these sources looks like FBM with

$$H = \frac{1}{2}\cdot(3-\min(\alpha_{\text{on}},\alpha_{\text{off}})).$$

The exact definition of this convergence is given in detail in [41] – it should be noted that both the number of sources is large and time is rescaled. This would suggest that the loss curve of a large number of these sources looks like $b^{2(1-H)}$. However, from Theorem 2.8 we know that it behaves as $(\alpha_{\text{on}}-1)\log b$. Apparently, the limits that are taken (large aggregation, large buffer, time rescaling) do *not* commute.

Notice also that in case of Exponential on-times and Pareto off-times, the aggregate still converges to FBM, whereas Theorem 2.7 gives that the overflow probability decays exponentially in the buffer size.

Claim 4.3 *If the on-times of the sources are heavy-tailed, so is the queue-length distribution.*

This claim needs to be stated a little more precisely. As shown by Dumas and Simonian [14], the overflow probability decays exponentially in the buffer size as long as the peak rates of the sources with heavy-tailed on-times *plus* the mean rates of the sources with Exponential on-times is below the link rate. If this is not the case, then the statement is formally true, in the sense that the overflow probability decays in a subexponential way in the buffer size.

However, as the experiments in Section 3 showed, in practical terms, in hardly any scenario the large buffer regime is reached; the small buffer regime seems to be more relevant as long as the on-times are not endowed with extremely heavy tails, the delay requirement is not extremely loose, and there is a reasonable level of aggregation.

Claim 4.4 *The loss probabilities in a multiplexing system are determined by the* tails *of the distributions of activity and silence periods of the sources.*

This 'myth' was already falsified in [17,18,34]: in 'realistic scenarios' there a critical time-scale was found beyond which the correlations do not significantly affect the overflow probability. In other words: Markovian models that capture the short-term correlations (up to the critical time-scale) are well-suited to predict the overflow probability. The exact shapes of the *tails* of the distributions of the on and off-times are therefore of minor importance. By 'realistic scenarios' we again mean that tails are not extremely fat, the delay requirement is somewhat stringent, and there is a fair amount of multiplexing. One could expect that in practical scenarios, the distribution 'at the left hand side' could be more relevant, i.e., the probability of extremely short on- and off-times. It could be seen easily that there could be important that with relatively high probability there is an extremely small interarrival times or silence periods (for a given mean).

Based on our objections to Claims 4.1 up to 4.4, clearly the statement 'Long range dependence leads to performance degradation' is not universally true.

5. Conclusions

Starting from the generic on-off source model, we have assessed the impact of long-range dependence (LRD) on queueing performance. Importantly, this impact is parametrized by the performance criterion imposed, as shown in detail in Section 3. If the delay requirement is 'tight', the number of admissible sources is insensitive in the distributions of the bursts and silences. The second relevant factor is the so-called 'level of aggregation': if the link rate is large compared to the peak rate of the source (which is not too large compared to the mean rate of the source), a fairly high utilization can be achieved, even when the delay requirements are tight. Hence, from a more practical point of view, the claim that LRD leads to performance degradation does certainly not hold in general. As illustrated in Section 4, there are also a number of theoretical objections to this statement.

Acknowledgement. We would like to thank Ward Whitt (AT&T Labs) and Alan Weiss (Bell Labs) for their valuable comments on previous drafts.

Appendix A. Computation of the loss curve and the acceptance curve

In this appendix we indicate how to compute the loss curve for the case that A and S are discrete random variables. In this case the distribution of the residual activity period $A^\star$ is given by

$$\mathbb{P}(A^\star > k) = \frac{1}{\mathbb{E}A} \sum_{l=k}^{\infty} \mathbb{P}(A > l).$$

A similar result holds for the residual silence distribution. Abbreviate

$$a_k := \mathbb{P}(A = k); \quad s_k := \mathbb{P}(S = k);$$
$$a_k^\star := \mathbb{P}(A^\star = k); \; s_k^\star := \mathbb{P}(S^\star = k).$$

Moment generating function. First we point how to compute moment generating function $\mathbb{E}\exp(\theta A(k))$. This can be done recursively, as follows. Clearly, in evident notation,

$$\mathbb{E}e^{\theta A(k)} = p\mathbb{E}_{A^\star} e^{\theta A(k)} + (1-p)\mathbb{E}_{S^\star} e^{\theta A(k)}.$$

Both terms can be evaluated as follows:

$$\mathbb{E}_{A^\star} e^{\theta A(k)} = \sum_{i=1}^{k-1} a_i^\star e^{\theta i} \mathbb{E}_S e^{\theta A(k-i)} + \sum_{i=k}^{\infty} a_i^\star e^{\theta k},$$
$$\mathbb{E}_{S^\star} e^{\theta A(k)} = \sum_{i=1}^{k-1} s_i^\star \mathbb{E}_A e^{\theta A(k-i)} + \sum_{i=k}^{\infty} s_i^\star,$$

where

$$\mathbb{E}_A e^{\theta A(j)} = \sum_{i=1}^{j-1} a_i e^{\theta i} \mathbb{E}_S e^{\theta B(j-i)} + \sum_{i=j}^{\infty} a_i e^{\theta j},$$
$$\mathbb{E}_S e^{\theta A(k)} = \sum_{i=1}^{j-1} s_i \mathbb{E}_A e^{\theta A(j-i)} + \sum_{i=j}^{\infty} s_i.$$

If the process alternates between two positive levels (rather than just on-off), it is convenient to write $A(k)$

as a on-off part $B(k)$ plus a part that is linear in k. This is done as follows. Let r_m denote the minimum rate, let r_p denote the peak rate, and define $r := r_p - r_m$. We can rewrite $A(k)$ as $r_m k + rB(k)$, where $B(k)$ is traffic generated by an on-off source with peak rate 1.

Loss curve and acceptance curve. When calculating $I(b)$, the variational problem

$$\inf_{k\in\mathbb{N}} \sup_{\theta} \left(\theta(b+ck) - \log \mathbb{E}e^{\theta A(k)}\right)$$

has to be solved. It is easy to find $\theta(k)$, i.e., the optimizing argument of the inner optimization for fixed k. This is because the function is convex in θ; there is a unique optimizer in $\mathbb{R}_+$. Then the infimum over k has to be computed – there is no nice concavity property, unfortunately.

When calculating $j(D)$ in (7), we lack the convexity property of the optimization over θ. However, the complexity of the numerical procedure turns out to be comparable to that of the loss curve.

The main effort in computing the acceptance curve numerically consists of computing the moment generating function $\mathbb{E}\exp(\theta A(k))$ for various combinations of θ and k. In order to compute this moment generating function for a given k, one has to compute for all $l = 1, \ldots, k-1$. It is not hard to see that hence the complexity of computing $\mathbb{E}\exp(\theta A(k))$ equals $O\left(\sum_{l=1}^{k} O(l)\right) = O(k^2)$. Call the optimizing k in (7) $k^\star$. Since for fixed D the maximum value of k is approximately $k^\star$, the complexity of computing $j(D)$ is roughly $O\left(k^{\star 2}\right)$.

Recall from Section 2.3 that for Weibull and Geometric on-times $k^\star$ grows linearly in D, and for Pareto on-times the growth of $k^\star$ is even superlinear in D. Thus the computing time for $j(D)$ increases rapidly for large D. For this reason we choose interrupt our calculations for D equal to some $k_{\max}$. We chose $k_{\max} = 1500$ in our numerical computations.

An approximation for the acceptance curve for higher delays can be obtained by increasing the packet size. Effectively, this redefines the time unit: the interarrival time of packets (within a burst) increases. In this way the rapid growth of $k^\star$ (as function of D) can be controlled.

Appendix B. Acceptance curve for small delays

In this appendix we derive a generic property of the acceptance curve. For small values of D we expect that the number of sources to be admitted grows rapidly, based on the square root in Theorem 2.5. Then, for small $b = cD$, we have to solve

$$J(D)\cdot\alpha\left(\frac{c}{J(D)}\right) + J(D)\cdot\beta\left(\frac{c}{J(D)}\right)\cdot\sqrt{\frac{b}{J(D)}} = \delta.$$

Let us try the approximation $J(D) \approx J(0) + K\sqrt{b}$ for some positive constant K. Abbreviate

$$\alpha_J := \alpha\left(\frac{c}{J(0)}\right);\ \alpha'_J := \alpha'\left(\frac{c}{J(0)}\right);$$
$$\beta_J := \beta\left(\frac{c}{J(0)}\right);\ \beta'_J := \beta'\left(\frac{c}{J(0)}\right).$$

Notice that $\delta = J(0)\alpha_J$ due to Theorem 2.5. We get, neglecting terms of order $O(b)$,

$$\begin{aligned}
\delta &= \left(J(0)+K\sqrt{b}\right)\cdot\alpha\left(\frac{c}{J(0)+K\sqrt{b}}\right) + \left(J(0)+K\sqrt{b}\right)\\
&\quad\cdot\beta\left(\frac{c}{J(0)+K\sqrt{b}}\right)\cdot\sqrt{\frac{b}{J(0)+K\sqrt{b}}}\\
&= \left(J(0)+K\sqrt{b}\right)\cdot\left(\alpha\left(\frac{c}{J(0)} - \frac{cK}{J^2(0)}\sqrt{b}\right)\right.\\
&\quad\left.+\beta\left(\frac{c}{J(0)} - \frac{cK}{J^2(0)}\sqrt{b}\right)\cdot\sqrt{\frac{b}{J(0)}}\right)\\
&= \left(J(0)+K\sqrt{b}\right)\cdot\left(\alpha_J - \frac{cK}{J^2(0)}\sqrt{b}\alpha'_J\right.\\
&\quad\left.+\left(\beta_J - \frac{cK}{J^2(0)}\sqrt{b}\beta'_J\right)\cdot\sqrt{\frac{b}{J(0)}}\right)\\
&= \delta + \sqrt{b}\left(K\alpha_J - \frac{cK}{J(0)}\alpha'_J + \sqrt{J(0)}\beta_J\right).
\end{aligned}$$

This gives us

$$\begin{aligned}
K &= \left(\alpha'_J\frac{c}{J(0)} - \alpha_J\right)^{-1}\left(\sqrt{J(0)}\beta_J\right)\\
&= \left(\log\left(\frac{1-p}{1-c/J(0)}\right)\right)^{-1}\left(\sqrt{J(0)}\beta_J\right).
\end{aligned}$$

As K is a finite positive number, our initial guess $J(D) \approx J(0) + K\sqrt{b}$ turns out to hold. Interestingly, the acceptance curve is insensitive in the higher moments of activities and silences (just like the loss curve is). This is an immediate consequence of the fact that $J(0)$ only depends on the on- and off-times through p, and β_J through $\mathbb{E}A$ and $\mathbb{E}S$. We notice that the acceptance curve grows rapidly for small b, namely like $\sqrt{b}$.

References

[1] Anick, D.; Mitra, D.; Sondhi, M.: Stochastic theory of a data-handling system with multiple sources. Bell System Tech. J. **61** (1982), 1871–1894.

[2] Asmussen, S.; Rubinstein, R.: Steady state rare event simulation in queueing models and its complexity properties. In Dshalalow, J. (Ed.), Advances in queueing theory, theory, methods and open problems, 429–461, Boca Raton, USA: CRC Press, 1995.

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

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

[5] Boxma, O.; Dumas, V.: Fluid queues with long-tailed activity period distributions. Comput. Commun. **21** (1998), 1509–1529.

[6] Choudhury, G.; Lucantoni, D.; Whitt, W.: Squeezing the most out of ATM. IEEE Trans. Commun. **44** (1996), 203–217.

[7] Cohen, J.: Some results on regular variation for distributions in queueing and fluctuation theory. J. Appl. Probability **10** (1973), 343–353.

[8] Courcoubetis, C.; Weber, R.: Buffer overflow asymptotics for a buffer handling many traffic sources. J. Appl. Probability **33** (1996), 886–903.

[9] Crovella, M.; Bestavros, A.: Self-similarity in World Wide Web traffic: evidence and possible causes. IEEE/ACM Trans. Networking **5** (1997), 835–846.

[10] Dembo, A.; Zeitouni, O.: Large Deviations Techniques and Applications. Jones and Bartlett, Boston, 1993.

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

[12] Duffield, N.; O'Connell, N.: Large deviations and overflow probabilities for the general single server queue, with applications. Proc. Cambridge Philos. Soc. **118** (1995), 363–374.

[13] Duffield, N.; Whitt, W.: Large deviations of inverse processes with nonlinear scalings. Ann. Appl. Probability **8** (1998), 995–1026.

[14] Dumas,V.; Simonian, A.: Asymptotic bounds for the fluid queue fed by subexponential on-off sources. Preprint.

[15] Elwalid, A.; Mitra, D.; Wentworth, R.: A new approach for allocating buffers and bandwidth to heterogeneous, regulated traffic in an ATM node. IEEE J. Selected Areas Commun. **13** (1995), 1115–1127.

[16] Feldmann. A.: Characteristics of TCP connection arrivals. Internal memorandum AT&T Labs, available at `http://www.research.att.com/~anja/feldmann/papers.html#traffic`

[17] Grossglauser, M.; Bolot, J.-C.: On the relevance of long-range dependence in network traffic. IEEE/ACM Trans. Networking **7** (1999), 629–640.

[18] Heyman, D.; Lakshman, T.: What are the implications of long-range dependence for VBR traffic engineering? IEEE/ACM Trans. Networking **4** (1996), 301–317.

[19] Heyman, D.; Lakshman, T.; Liu, D.: Assessing the effects of short-range and long-range dependence on overflow probabilities. Proceedings 13th ITC specialist seminar: IP traffic measurement, modeling, and management, 19.1–19.11, 2000.

[20] Jelenković, P.; Lazar, A.: Asymptotic results for multiplexing on-off sources with subexponential on-times. Advances in Applied Probability **31** (1999), 394–421.

[21] Jelenković, P.; Lazar, A.; Semret, N.: The effect of multiple time scales and subexponentiality in MPEG video streams on queueing behavior. IEEE J. Selected Areas Commun. **15** (1997), 1052–1071.

[22] Kosten, L.: Stochastic theory of data-handling systems with groups of multiple sources. In H. Rudin and W. Bux (Eds.), Performance of Comput.-Communication Systems, 321–331, Elsevier, Amsterdam, 1984.

[23] Leland, W.; Taqqu, M.; Willinger, W.; Wilson, D.: On the self-similar nature of Ethernet traffic. IEEE/ACM Trans. Networking **2** (1994), 1–15.

[24] Likhanov, N.; Mazumdar, R.: Cell loss asymptotics in buffers fed with a large number of independent stationary sources. Proc. IEEE Infocom 1998, 339–346.

[25] Mandjes, M.; Borst, S.: Overflow behavior in queues with many long-tailed inputs. Adv. Appl. Probability **32** (2000), 1150–1167.

[26] Mandjes, M.; Kim, J.H.: Large deviations for small buffers: an insensitivity result. Queueing Systems **33** (2001), 349–362.

[27] Mandjes, M.; Mitra, D.; Scheinhardt, W.: A simple model of network access: feedback adaptation of rates and admission control. Comput. Networks **41** (2003), 489–504.

[28] Mandjes, M.; Ridder, A.: Finding the conjugate of Markov fluid processes. Probability Eng. Informational Sci. **9** (1995), 297–315.

[29] Massoulié, L.; Simonian, A.: Large buffer asymptotics for the queue with FBM input. J. Appl. Probability **36** (1999), 894–906.

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

[31] Pakes, A.: On the tail of waiting time distributions. J. Appl. Probability **12** (1975), 555–564.

[32] Paxson, V.; Floyd, S.: Wide area traffic: the failure of Poisson modeling. IEEE/ACM Trans. Networking **3** (1995), 226–244.

[33] Rose, O.: Traffic modeling of variable bit rate MPEG video and its impact on ATM networks. Ph.D. thesis, University of Würzburg, Germany, 1997.

[34] Ryu, B.; Elwalid, A.: The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities. Comput. Commun. Rev. **26** (1996), 3–14.

[35] Shwartz, A.; Weiss, A.: Large deviations for performance analysis, queues, communication, and computing. Chapman and Hall, New York, 1995.

[36] Simonian, A.; Guibert, J.: Large deviations approximation for fluid queues fed by a large number of on/off sources. IEEE J. Selected Areas Commun. **13** (1995), 1017–1027.

[37] Sriram, K.; Whitt, W.: Characterizing superposition arrival processes in packet multiplexers for voice and data. IEEE J. Selected Areas Commun. **4** (1986), 833–846.

[38] van der Wal, K.; Mandjes, M.; Bastiaansen, H.: Delay performance analysis of the new internet services with guaranteed QoS. Proceedings IEEE **85** (1997), 1947–1957.

[39] Weiss, A.: A new technique of analyzing large traffic systems. Adv. Appl. Probability **18** (1986), 506–532.

[40] Whitt, W.: Tail probabilities with statistical multiplexing and effective bandwidth in multi-class queues. Telecommunication Systems **2** (1994), 71–107.

[41] Willinger, W.; Taqqu, M.; Sherman, R.; Wilson, D.: Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. Comput. Commun. Rev. **25** (1995), 100–113.

[42] Wischik, D.: Sample path large deviations for queues with many inputs. Ann. of Applied Probability **11** (2001), 379–404.

M. Mandjes received M.Sc. degrees in mathematics and econometrics from the Vrije Universiteit (Free University), Amsterdam, both cum laude. In 1996 he received his Ph.D. degree (thesis: Rare event analysis of communication networks), again.from the Free University, Amsterdam. Autumn 1996 he joined KPN Research (currently TNO Telecom), where he mainly worked on ATM and IP performance analysis in an operational context.
He was involved in a number of European research consortia, e.g. COST 257 and CAShMAN. From 1999 till 2001, he was employed at Bell Laboratories/Lucent Technologies, Murray Hill NJ, USA, where he worked as a member of technical staff in the Mathematics of Networks and Systems group.
From October 2000, he has a joint position as a senior researcher at CWI (Center for Mathematics and Computer Science, Amsterdam, the Netherlands), and as a full professor of Stochastic Operations Research at the University of Twente (Faculty of Electrical Engineering, Mathematics, and Computer Science, Enschede, the Netherlands).

N.K. Boots received M.Sc. degrees in mathematics and econometrics from the Vrije Universiteit (Free University), Amsterdam.
In 2002 he received his Ph.D. degree (thesis: Rare event simulation in models with heavy-tailed random variables), again from the Free University, Amsterdam. He is currently working for ABN AMRO Bank, Amsterdam, the Netherlands, as a member of the department Market Risk – Modelling & Product Analysis, where he works as a derivatives researcher.