Subexponential loss rates in a GI/GI/1 queue with applications

Predrag R. Jelenković

Department of Electrical Engineering, Columbia University, New York, NY 10027, USA

Received April 1998; revised February 1999

Consider a single server queue with i.i.d. arrival and service processes, $\{A,\ A_n, n \geqslant 0\}$ and $\{C,\ C_n, n \geqslant 0\}$, respectively, and a finite buffer B. The queue content process $\{Q_n^B,\ n \geqslant 0\}$ is recursively defined as $Q_{n+1}^B = \min((Q_n^B + A_{n+1} - C_{n+1})^+, B)$, $q^+ = \max(0, q)$. When $\mathbb{E}(A - C) < 0$, and A has a subexponential distribution, we show that the stationary expected loss rate for this queue $\mathbb{E}(Q_n^B + A_{n+1} - C_{n+1} - B)^+$ has the following explicit asymptotic characterization:

$$\mathbb{E}(Q_n^B + A_{n+1} - C_{n+1} - B)^+ \sim \mathbb{E}(A - B)^+ \quad \text{as } B \to \infty,$$

independently of the server process C_n. For a fluid queue with capacity c, M/G/∞ arrival process A_t, characterized by intermediately regularly varying *on* periods τ^{on}, which arrive with Poisson rate Λ, the average loss rate $\lambda_{\mathrm{loss}}^B$ satisfies

$$\lambda_{\mathrm{loss}}^B \sim \Lambda\ \mathbb{E}(\tau^{\mathrm{on}}\eta - B)^+ \quad \text{as } B \to \infty,$$

where $\eta = r + \rho - c$, $\rho = \mathbb{E}A_t < c$; r $(c \leqslant r)$ is the rate at which the fluid is arriving during an *on* period. Accuracy of the above asymptotic relations is verified with extensive numerical and simulation experiments. These explicit formulas have potential application in designing communication networks that will carry traffic with long-tailed characteristics, e.g., Internet data services.

Keywords: long-tailed traffic models, subexponential distributions, long-range dependency, network multiplexer, finite buffer queue, fluid flow queue, M/G/∞ process

1. Introduction

An increasing body of the literature on statistical data analysis has demonstrated the presence of long-tailed (subexponential) characteristics in communication network traffic streams. Early discoveries on the long-tailed nature of Ethernet traffic was reported in [29]. Long-tailed characteristics of the scene length distribution of MPEG video streams were explored in [22,26]. The implications of transporting Internet data applications over the traditional Public Switched Telephone Network were investigated in [19].

These empirical findings have encouraged theoretical developments in the modeling and analysis of long-tailed (heavy-tailed) phenomena. In this area there have been two basic approaches: self-similar processes and fluid renewal models with long-tailed renewal distributions. The investigation of queueing systems with self-similar long-range dependent arrival processes can be found in [14,15,30,34,36,38,41,42].

In this paper we focus on fluid renewal models. Basic tools for the analysis of these types of models with a single long-tailed arrival stream are the classical results on subexponential asymptotic behavior of the waiting time distribution in a GI/GI/1 queue [12,35,43] (these results were used in [2,21]). Asymptotic expansion refinements of these results can be found in [1,44]. Generalizations to queueing processes (random walks) with dependent increments were investigated in [4,5,24].

Queueing models with multiple long-tailed arrival streams are of particular interest for engineering communication networks. Unfortunately, the analysis of these models is much more difficult due to the complex dependency structure in the aggregate arrival process [21]. An intermediate case of multiplexing a single long-tailed stream with exponential streams was investigated in [8,25,40].

For the case of multiplexing more than two long-tailed arrival processes general bounds were obtained in [10,32]. In [8] a limiting process obtained by multiplexing an infinite number of on–off sources with regularly varying on periods was analyzed. This limiting arrival process, the so-called $M/G/\infty$ process [37], appears to be quite promising for the analysis. In [25] an explicit asymptotic formula for the behavior of the infinite buffer queue length distribution with $M/G/\infty$ arrivals was derived. In the same paper it was shown with simulation experiments that the derived asymptotic formula yields a good approximation for multiplexing finitely many long-tailed on–off sources. An asymptotic expression for the expected value of the first passage time in a fluid queue with $M/G/\infty$ arrivals was derived in [20]. A recent survey of results on fluid queues with long-tailed arrival processes can be found in [9].

All of the previously mentioned results in the literature on stationary queueing analysis assume an infinite buffer queue. This assumption is applicable to queueing systems that are designed with very large buffers such that the losses are essentially zero. The queue length distribution can be used in this situation as an indication of the delay experienced in the system. However, in engineering network switches it is very common to design them as loss systems. The main performance measures for these systems are loss probabilities and loss rates. Obtaining asymptotic approximations for this performance measures under the assumption of subexponential arrival sequences is the primary motivation for the investigation of this paper.

The main contributions of this paper, presented in theorems 4 and 5, are explicit asymptotic characterizations of the loss rates in finite buffer queues with subexponential arrival sequences. Theorem 5, in combination with the results from [25,39], yields a straightforward asymptotic formula for the loss rate in a fluid queue with long-tailed $M/G/\infty$ arrivals (see theorem 7). Accuracy of the theoretical asymptotic results is demonstrated with many numerical and simulation experiments. We believe that the exactness and explicit nature of the derived approximate expressions will make them useful tools in designing efficient and reliable network switches.

The rest of the paper is organized as follows. First, in section 2, we present a preliminary discussion of subexponential distributions and infinite buffer queueing analysis. Section 3 investigates a discrete time finite buffer queue. The main results are summarized in theorems 4 and 5. The fluid aspect of the problem is explored in section 4, theorems 6 and 7. Numerical and simulation examples that illustrate the efficacy of our approximations are contained in section 5. The paper is concluded in section 6. To simplify the reading process, the majority of the proofs are given in the appendix.

2. Subexponential distributions and GI/GI/1 queueing analysis

This section presents a preliminary discussion on the long-tailed and subexponential distributions and the asymptotic analysis of an infinite buffer queue under the assumption of subexponentiality.

Let A be a nonnegative random variable with a finite mean and $\mathbb{P}[A > x] > 0$, $x \geqslant 0$. We say that A (or its distribution function (d.f.)) is *long-tailed* ($A \in \mathcal{L}$) if

$$\lim_{x\to\infty} \frac{\mathbb{P}[A > x + y]}{\mathbb{P}[A > x]} = 1, \quad y \geqslant 0. \tag{1}$$

A (or its d.f.) is said to be *subexponential* ($A \in \mathcal{S}$) if

$$\lim_{x\to\infty} \frac{\mathbb{P}[A_1 + \cdots + A_n > x]}{\mathbb{P}[A > x]} = n, \tag{2}$$

where A_n, $n \geqslant 1$, is a sequence of independent copies of A. The following subclass of subexponential random variables was introduced in [27]. A random variable A with a finite mean (or its d.f.) is said to be in $\mathcal{S}^*$ if

$$\lim_{x\to\infty} \int_0^x \frac{\mathbb{P}[A > x - y]}{\mathbb{P}[A > x]} \mathbb{P}[A > y]\,\mathrm{d}y = 2\mathbb{E}A. \tag{3}$$

This class of subexponential distribution is closed under the tail integration, i.e., if A_e is the remaining life time random variable of A defined as $\mathbb{P}[A_e \leqslant x] = (\int_0^x \mathbb{P}[A > u]\,\mathrm{d}u)/\mathbb{E}A$, $x \geqslant 0$, then $A \in \mathcal{S}^*$ implies $A_e \in \mathcal{S}^*$. A general relationship between the previous three classes of long-tailed random variables is $\mathcal{S}^* \subset \mathcal{S} \subset \mathcal{L}$. For a brief introduction to long-tailed and subexponential distributions the reader is referred to the Appendix of this special issue of Queueing Systems. A recent survey on subexponential distributions can be found in [18]. Well known examples of subexponential distributions incorporate regularly varying distributions (in particular, Pareto), some Weibull and Log-normal distributions.

Subexponential random variables have played a role in queueing theory since the classical results of Cohen [12] and Pakes [35] on the asymptotic behavior of the waiting time process in a GI/GI/1 queue. Here, we give a formal definition of a GI/GI/1 queue waiting time process. Let $\{A,\ A_n,\ n \geqslant 1\}$ and $\{C,\ C_n,\ n \geqslant 1\}$ be two independent sequences of i.i.d. random variables. Then for any initial condition Q_0 the queueing process $\{Q_n,\ n \geqslant 0\}$ is uniquely defined by the following (Lindley's) recursion:

$$Q_{n+1} = (Q_n + A_{n+1} - C_{n+1})^+, \quad n \geqslant 0, \tag{4}$$

where $q^+ = \max(0, q)$. This recursion has several possible interpretations. If one assumes that A_n, $n \geqslant 1$, are customers' service requirements and C_n, $n \geqslant 1$, are their inter-arrival times, then Q_n represents the waiting time process in a GI/GI/1 queue. If one thinks of (4) as being an infinite buffer discrete time queue with A_n representing the amount of work that arrives at time n and C_n the amount of work that is served at time n, then Q_n represents the queue length process for this queue. We will simply refer to Q_n, $n \geqslant 0$, as the infinite buffer queueing process and to A_n and C_n as the arrival and service processes, respectively.

According to the classical result of Loynes [31], under the stability condition $\mathbb{E}A_n < \mathbb{E}C_n$, this recursion admits a unique stationary solution, and for all initial conditions $\mathbb{P}[Q_n \leqslant x]$ converges to the stationary distribution $\mathbb{P}[Q \leqslant x]$. For the rest of this paper, unless otherwise indicated, we will assume that all queueing systems under consideration are in their stationary regimes.

Often, it is easier to conduct numerical computations with lattice valued random variables than with continuous ones. In this context one may be interested in computing queue occupancy probabilities. The following result gives an approximation of these probabilities under the subexponential assumption on the arrival sequence. Note that the result does not follow directly from Pakes' result [35], since the asymptotic behavior of $\mathbb{P}[Q > i]$ does not imply the asymptotics of $\mathbb{P}[Q = i]$. Throughout the paper, for any two real functions $f(x)$ and $g(x)$, we use the standard notation $f(x) \sim g(x)$ as $x \to \infty$ to denote $\lim_{x\to\infty} f(x)/g(x) = 1$, or equivalently, $f(x) = g(x)(1 + o(1))$ as $x \to \infty$.

Theorem 1. If A, C are integer valued, $A \in \mathcal{S}^*$, $\mathbb{P}[C \leqslant c] = 1$, $c < \infty$, and $\mathbb{E}A < \mathbb{E}C$, then the queue occupancy probabilities satisfy

$$\mathbb{P}[Q = i] \sim \frac{1}{\mathbb{E}C - \mathbb{E}A}\,\mathbb{P}[A > i] \quad \text{as } i \to \infty.$$

Proof. Given in appendix A.1. □

If one is only interested in the tail of $\mathbb{P}[Q > x]$, then under more general assumptions Pakes [35] has derived the following result.

Theorem 2. If $A_e \in \mathcal{S}$ and $\mathbb{E}A < \mathbb{E}C$, then the tail of the queue length-distribution satisfies

$$\mathbb{P}[Q > x] \sim \frac{\mathbb{E}A}{\mathbb{E}C - \mathbb{E}A}\,\mathbb{P}[A_e > x] \quad \text{as } x \to \infty.$$

3. Finite buffer queue

In this section we present our results on a discrete time finite buffer queue. The results are stated in theorems 3–5.

Let $\{A,\ A_n,\ n \geqslant 1\}$ and $\{C,\ C_n,\ n \geqslant 1\}$, as in the previous section, be two independent sequences of i.i.d. random variables. The evolution of a finite buffer queue is defined with the following recursion:

$$Q^B_{n+1} = \min\bigl(\bigl(Q^B_n + A_{n+1} - C_{n+1}\bigr)^+, B\bigr), \quad n \geqslant 0, \tag{5}$$

where B is the buffer size. It is clear that Q^B_n is a discrete time Markov process with state space $[0, B]$. By excluding a trivial situation of $C_n \equiv A_n$, i.e., by assuming that $\mathbb{P}[A_n = C_n] < 1$ in [13, chapter III.4] it was shown that this Markov process has a unique stationary distribution, and that for all initial conditions Q^B_0, Q^B_n converges to that stationary distribution. Unless otherwise indicated we will assume that the recursion (5) is in its stationary regime. Similarly as in (4), Q^B_ns can be interpreted as the uniformly bounded customer waiting times in a GI/GI/1 queue (see [13, chapter III.4]).

In the proofs of lemmas 1 and 2 and theorem 3 we will restrict our attention to A_n and C_n being lattice valued. Without loss of generality we can assume that A_n and C_n are integer valued. Next, denote the corresponding probabilities with $a_i = \mathbb{P}[A_n = i]$, $c_i = \mathbb{P}[C_n = i]$, and $x_i = \mathbb{P}[X_n = i]$, $i \geqslant 0$, where $X_n \stackrel{\text{def}}{=} A_n - C_n$. In addition, assume that C_n has a bounded support $\mathbb{P}[C_n \leqslant c] = 1$. Let

$$a(z) = \sum_{i=0}^{\infty} a_i z^i, \quad c(z) = \sum_{i=0}^{c} c_i z^i \quad \text{and} \quad x(z) = \sum_{i=-c}^{\infty} x_i z^i$$

be the probability generating functions (pgf) for A_n, C_n and X_n, respectively. It is easy to show that the stationary queue occupancy pgf

$$q^B(z) = \sum_{i=0}^{B} q^B_i z^i, \quad q^B_i = \mathbb{P}[Q^B_n = i], \quad 0 \leqslant i \leqslant B,\ B \in \mathbb{N}_0,$$

is equal to

$$q^B(z) = \frac{\sum_{i=0}^{c-1} \sum_{k=0}^{i} q^B_k x_{i-k-c}(z^c - z^i) + R^B(z)}{z^c - z^c x(z)}, \tag{6}$$

where

$$R^B(z) \stackrel{\text{def}}{=} \sum_{i=B+1}^{\infty} \sum_{k=0}^{B} q^B_k x_{i-k}\bigl(z^{B+c} - z^{i+c}\bigr). \tag{7}$$

In order to prove our main results we need the following two technical lemmas. Let $\nu^B_i \stackrel{\text{def}}{=} q^B_i / q^B_0$, $\infty \geqslant B \geqslant i \geqslant 0$ $(\nu^B_0 = 1)$.

Lemma 1. If $\mathbb{E}A < \mathbb{E}C$, and $\mathbb{P}[C \leqslant c] = 1, c < \infty$, then there exists a positive constant K_1, such that for all $B \geqslant 0$

$$\left|\nu_i^B - \nu_i^\infty\right| \leqslant K_1 \mathbb{P}\left[Q_n^B + A_{n+1} - C_{n+1} > B\right]\delta^{B-i}, \quad 0 \leqslant i \leqslant B,$$

where $\delta = 0$ if $c = 1$, and $\delta < 1$ if $c > 1$.

Remark. Note that this lemma does not require subexponentiality of A.

Proof. Let us first prove the case $c = 1$. Observe that q_i^B, $0 \leqslant i \leqslant B$, satisfy the following set of B independent equations:

$$\begin{aligned} q_0^B &= q_0^B x_0 + q_1^B x_{-1}, \\ q_1^B &= q_0^B x_1 + q_1^B x_0 + q_2^B x_{-1}, \\ &\vdots \\ q_{B-1}^B &= q_0^B x_{B-1} + \cdots + q_B^B x_{-1}. \end{aligned}$$

Since $\mathbb{E}A < \mathbb{E}C = c_1 \leqslant 1 \Rightarrow x_{-1} = a_0 c_1 > 0$, we see that ν_i^B is uniquely defined by the preceding set of equations. Similarly, q_i^∞, $B \geqslant i \geqslant 0$, satisfy exactly the same set of equations and therefore $\nu_i^B = \nu_i^\infty$, $B \geqslant i \geqslant 0$. This proves the case $c = 1$.

The case $c > 1$ is much more involved and is presented in appendix A.2. □

In order to make the preceding lemma useful we need the following bound on the buffer overflow probability.

Lemma 2. If $A_e \in \mathcal{S}$, $\mathbb{E}A < \mathbb{E}C$, and $\mathbb{P}[C \leqslant c] = 1$, $c < \infty$, then

$$\mathbb{P}\left[Q_n^B + A_{n+1} - C_{n+1} > B\right] = \mathrm{o}\left(\mathbb{P}[A_e > B]\right) \quad \text{as } B \to \infty.$$

(Recall that $\mathbb{P}[A_e \leqslant x] = \int_0^x \mathbb{P}[A > u]\,\mathrm{d}u/\mathbb{E}A$.)

Proof. Given in appendix A.3. □

Theorem 3. If $A_e \in \mathcal{S}$, $\mathbb{E}A < \mathbb{E}C$, and $\mathbb{P}[C \leqslant c] = 1$, $c < \infty$, then

$$\begin{aligned} 1 - \frac{q_0^\infty}{q_0^B} &= \mathbb{P}\left[Q^\infty > B\right]\left(1 + \mathrm{o}(1)\right) = \frac{\mathbb{E}A}{\mathbb{E}C - \mathbb{E}A}\mathbb{P}[A_e > B]\left(1 + \mathrm{o}(1)\right) \\ &= \frac{\mathbb{E}(A - B)^+}{\mathbb{E}C - \mathbb{E}A}\left(1 + \mathrm{o}(1)\right) \end{aligned}$$

as $B \to \infty$.

Remark. Observe that in the case $\mathbb{P}[C \leqslant 1] = 1$, lemma 1 implies the following identity:

$$q_i^B = \frac{q_i^\infty}{\mathbb{P}[0 \leqslant Q^\infty \leqslant B]}. \tag{8}$$

Thus, the theorem follows directly from theorem 2, i.e., lemma 2 is not needed. A similar identity exists when C is exponentially distributed (see [45, eq. (3.6)]).

Proof. By combining lemmas 1 and 2 we compute

$$\sum_{i=0}^{B} \nu_i^B \leqslant \sum_{i=0}^{B} \nu_i^\infty + K_1 \sum_{i=0}^{B} \delta^{B-i} \mathrm{o}\big(\mathbb{P}[A_e > B]\big)$$
$$\leqslant \sum_{i=0}^{B} \nu_i^\infty + \frac{K_1}{(1-\delta)} \mathrm{o}\big(\mathbb{P}[A_e > B]\big),$$

which is equivalent to

$$1 - \frac{q_0^\infty}{q_0^B} \geqslant \mathbb{P}\big[Q^\infty > B\big] - \mathrm{o}\big(\mathbb{P}[A_e > B]\big). \tag{9}$$

Application of theorem 2 shows that $\mathbb{P}[A_e > B]$ is asymptotically proportional to $\mathbb{P}[Q^\infty > B]$, which, when replaced in (9), yields the lower bound

$$1 - \frac{q_0^\infty}{q_0^B} \geqslant \mathbb{P}\big[Q^\infty > B\big]\big(1 + \mathrm{o}(1)\big) = \frac{\mathbb{E}A}{\mathbb{E}C - \mathbb{E}A} \mathbb{P}[A_e > B]\big(1 + \mathrm{o}(1)\big) \quad \text{as } B \to \infty.$$

The upper bound can be proved in exactly the same manner. We omit the details. This proves the first two equalities of the theorem. The third equality follows from

$$\mathbb{E}(A - B)^+ = \int_0^\infty \mathbb{P}\big[(A - B)^+ > x\big]\,\mathrm{d}x = \int_B^\infty \mathbb{P}[A > x]\,\mathrm{d}x. \tag{10}$$

This finishes the proof of the theorem. □

Theorem 4. If $A_e \in \mathcal{S}$ and $\mathbb{E}A < \mathbb{E}C$, then the stationary loss rate $\mathbb{E}(Q_n^B + A_{n+1} - C_{n+1} - B)^+$ satisfies

$$\mathbb{E}\big(Q_n^B + A_{n+1} - C_{n+1} - B\big)^+ = \mathbb{E}(A - B)^+\big(1 + \mathrm{o}(1)\big) \quad \text{as } B \to \infty.$$

Remarks.

(i) In this theorem we do not assume that A and C are lattice valued.

(ii) From the theorem we can derive the probability that the work is lost

$$p(B) \overset{\text{def}}{=} \mathbb{E}\big(Q_n^B + A_{n+1} - C_{n+1} - B\big)^+ / \mathbb{E}A,$$

which in conjunction with (10) can be expressed in the following compact form

$$p(B) \sim \mathbb{P}[A_e > B] \quad \text{as } B \to \infty.$$

(iii) This theorem is an improvement of a theorem from the original version of the paper [23] which was proved under the assumption of A being regularly varying $\mathbb{P}[A > x] = l(x)/x^\alpha$ with index $\alpha > 2$.

Proof. Assume first that A_n and C_n are integer valued and that $\mathbb{P}[C_n \leqslant c] = 1$, $c < \infty$. From lemma 1 it follows that for any fixed i

$$q_i^B \frac{q_0^\infty}{q_0^B} = q_i^\infty + \mathrm{O}\bigl(\delta^B\bigr). \tag{11}$$

Next, by using the flow conservation law in the queue steady state regime (loss rate = arrival rate − departure rate) we compute

$$\begin{aligned}
&\mathbb{E}\bigl(Q_n^B + A_{n+1} - C_{n+1} - B\bigr)^+ \\
&\quad = \mathbb{E}A - \sum_{i=0}^{c} \mathbb{P}[C_{n+1} = i]\left(i\mathbb{P}\bigl[Q_n^B + A_{n+1} > i\bigr] + \sum_{k=0}^{i} k\mathbb{P}\bigl[Q_n^B + A_{n+1} = k\bigr]\right) \\
&\quad = \mathbb{E}A - \mathbb{E}C + \sum_{i=0}^{c}\sum_{k=0}^{i} (i-k)\mathbb{P}[C_{n+1} = i]\mathbb{P}\bigl[Q_n^B + A_{n+1} = k\bigr] \\
&\quad = \mathbb{E}A - \mathbb{E}C + \sum_{i=0}^{c}\sum_{k=0}^{i}\sum_{j=0}^{k} (i-k)\mathbb{P}[C_{n+1} = i]\mathbb{P}[A_{n+1} = k-j]q_j^B
\end{aligned}$$

which in conjunction with (11) yields

$$\begin{aligned}
&\mathbb{E}\bigl(Q_n^B + A_{n+1} - C_{n+1} - B\bigr)^+ \\
&\quad = \mathbb{E}A - \mathbb{E}C + \frac{q_0^B}{q_0^\infty}\sum_{i=0}^{c}\sum_{k=0}^{i} (i-k)\mathbb{P}[C_{n+1} = i]\mathbb{P}\bigl[Q_n^\infty + A_{n+1} = k\bigr] + \mathrm{O}\bigl(\delta^B\bigr).
\end{aligned}$$

Combining the expression above with a similar expression for $B = \infty$ (loss = 0), we arrive at

$$\mathbb{E}\bigl(Q_n^B + A_{n+1} - C_{n+1} - B\bigr)^+ = (\mathbb{E}C - \mathbb{E}A)\left(\frac{q_0^B}{q_0^\infty} - 1\right) + \mathrm{O}\bigl(\delta^B\bigr),$$

which together with theorem 3 and $\delta^B = \mathrm{o}(\mathbb{P}[A_e > B])$ as $B \to \infty$ completes the proof of the theorem for the case A_n and C_n being integer valued and C_n being bounded.

In general, we can easily obtain a lower bound

$$\begin{aligned}
\mathbb{E}\bigl(Q_n^B + A_{n+1} - C_{n+1} - B\bigr)^+ &\geqslant \mathbb{E}(A_{n+1} - C_{n+1} - B)^+ \\
&= \int_B^\infty \mathbb{P}[A_{n+1} - C_{n+1} > x]\,\mathrm{d}x \\
&\sim \mathbb{E}(A - B)^+ \quad \text{as } B \to \infty;
\end{aligned} \tag{12}$$

for the last asymptotic relation we have used $A_e \in \mathcal{L}$. If A_n and C_n are integer valued and C_n is unbounded we can always choose a truncated service variable $C_n^c = \min(C_n, c)$, with c being sufficiently large such that $\mathbb{E}A_n < \mathbb{E}C_n^c$. Let $Q_n^{B,c}$

be the queueing process that corresponds to the arrival process A_n and a modified service process C_n^c. It is clear that $Q_n^{B,c}$ is stochastically larger than Q_n^B, and that the corresponding loss rates satisfy

$$\begin{aligned}\mathbb{E}\big(Q_n^B + A_{n+1} - C_{n+1} - B\big)^+ &\leqslant \mathbb{E}\big(Q_n^{B,c} + A_{n+1} - C_{n+1}^c - B\big)^+ \\ &\sim \mathbb{E}(A-B)^+ \quad \text{as } B \to \infty. \end{aligned} \tag{13}$$

Now, (12) and (13) imply the conclusion of the theorem for A and C being integer, or in general lattice valued.

When A and C are non lattice, we can approximate them with lattice valued random variables A' and C' in the following way. First, for any $\Delta > 0$ such that $\mathbb{E}C - \mathbb{E}A + 2\Delta < 0$, we define the d.f.s for A' and C' as

$$\begin{aligned}\mathbb{P}\big[C' = \Delta i\big] &= \mathbb{P}\big[\Delta i \leqslant C < \Delta(i+1)\big], \quad i \geqslant 0, \\ \mathbb{P}\big[A' = \Delta i\big] &= \mathbb{P}\big[\Delta(i-1) \leqslant A < \Delta i\big], \quad i \geqslant 1.\end{aligned}$$

From these definitions it easily follows that for all $x \geqslant 0$

$$\begin{aligned}\mathbb{P}[C > x + \Delta] \leqslant \mathbb{P}\big[C' > x\big] &\leqslant \mathbb{P}[C > x], \\ \mathbb{P}[A > x] \leqslant \mathbb{P}\big[A' > x\big] &\leqslant \mathbb{P}[A > x - \Delta],\end{aligned}$$

which implies that $A' - C'$ is stochastically larger than $A - C$, $\mathbb{E}A' \leqslant \mathbb{E}A + \Delta < \mathbb{E}C - \Delta \leqslant \mathbb{E}C'$ and

$$\int_B^\infty \mathbb{P}\big[A' > u\big]\,\mathrm{d}u \sim \int_B^\infty \mathbb{P}[A > u]\,\mathrm{d}u \quad \text{as } B \to \infty.$$

Next, let $\{A_n',\ n \geqslant 1\}$ and $\{C_n',\ n \geqslant 1\}$ be two independent i.i.d. sequences whose d.f.s are equal to the d.f.s of A' and C', respectively, and consider a queue with buffer B which corresponds to sequences A_n' and C_n'. From the preceding discussion, the losses in this newly constructed queue are larger than the losses in the original queue and are asymptotically proportional to $\mathbb{E}(A_n' - B)^+ \sim \mathbb{E}(A_n - B)$ as $B \to \infty$. Hence, this yields an upper bound which in combination with the lower bound in (12) completes the proof. □

The following recursion, similar to the one in (5), will be useful in analyzing fluid queues in the following section:

$$W_{n+1}^B = \big(\min\big(W_n^B + A_{n+1}, B\big) - C_{n+1}\big)^+, \quad n \geqslant 0. \tag{14}$$

Under the same non-triviality condition as in the discussion of recursion (5), in [13, chapter III.5] it was shown that the Markov process W_n^B has a unique stationary distribution, and that for any initial condition W_0^B, W_n^B converges to that stationary distribution. Again, we assume that (14) operates in its stationary regime. Historically, recursion (14) has been studied in the context of finite dams (see in [13, chapter III.5]).

The next theorem shows that the loss rates for both queues (5) and (14) are asymptotically equivalent.

Theorem 5. If $A_e \in \mathcal{S}$, $\mathbb{E}A < \mathbb{E}C$, then

$$\mathbb{E}\big(W_n^B + A_{n+1} - B\big)^+ = \mathbb{E}(A-B)^+\big(1+\mathrm{o}(1)\big) \quad \text{as } B \to \infty.$$

Remark. This theorem generalizes a result from [45] which is true for A being regularly varying, and also for the case of C being exponential and A subexponential.

Proof. The proof of the lower bound is immediate:

$$\mathbb{E}\big(W_n^B + A_{n+1} - B\big)^+ \geqslant \mathbb{E}(A_{n+1} - B)^+. \tag{15}$$

In order to prove the upper bound, by using a simple inductive argument, one can show that for the same initial condition $W_0^B = Q_0^B$, W_n^B is bounded by Q_n^B. Hence, in stationarity

$$\mathbb{E}\big(W_n^B + A_{n+1} - B\big)^+ \leqslant \mathbb{E}\big(Q_n^B + A_{n+1} - B\big)^+. \tag{16}$$

Next, assume that C_n is bounded, i.e., $\mathbb{P}[C_n \leqslant c] = 1$. Then,

$$\mathbb{E}\big(W_n^B + A_{n+1} - B\big)^+ \leqslant \mathbb{E}\big(Q_n^B + A_{n+1} - C_{n+1} - B + c\big)^+, \tag{17}$$

which by theorem 4 implies

$$\begin{aligned}\mathbb{E}\big(W_n^B + A_{n+1} - B\big)^+ &\leqslant \big(1+\mathrm{o}(1)\big)\mathbb{E}(A_{n+1} - B + c)^+ \\ &= \big(1+\mathrm{o}(1)\big)\mathbb{E}(A_{n+1} - B)^+ \quad \text{as } B \to \infty,\end{aligned} \tag{18}$$

where the last asymptotic relation follows from $\mathbb{E}(A_{n+1} - B + c)^+ = \int_{B-c}^{\infty} \mathbb{P}[A > x]\,\mathrm{d}x$ and A_e being long-tailed.

For the case when C_n is not bounded we can always choose, similarly as in the proof of theorem 4, a truncated service variable $C_n^c = \min(C_n, c)$, with c being sufficiently large such that $\mathbb{E}A_n < \mathbb{E}C_n^c$. Let $W_n^{B,c}$ be the queueing process that corresponds to the arrival process A_n and a modified service process C_n^c. It is clear that $W_n^{B,c}$ is stochastically larger than W_n^B and that the corresponding loss rates satisfy

$$\begin{aligned}\mathbb{E}\big(W_n^B + A_{n+1} - B\big)^+ &\leqslant \mathbb{E}\big(W_n^{B,c} + A_{n+1} - B\big)^+ \\ &\sim \mathbb{E}(A-B)^+ \quad \text{as } B \to \infty.\end{aligned} \tag{19}$$

Thus, (15) and (19) finish the proof of the theorem. □

4. Finite buffer fluid queue

This section contains our results on fluid queues with finite buffers and long-tailed arrivals. In theorem 6 we obtain the asymptotic characterization of the loss rate of a fluid queue with a single on–off arrival process. An explicit asymptotic formula for the loss rate of a fluid queue with M/G/∞ arrival sequences is presented in theorem 7. This result is of special interest for designing communication network switches because the M/G/∞ arrival process represents a good aggregate model for multiplexing a large number of on–off sources (see [25]).

The physical interpretation of a fluid queue is that at any moment of time t, fluid is arriving to the system with rate a_t and is leaving the system with rate c_t. We term a_t and c_t to be the arrival and service processes, respectively. The evolution of the amount of fluid in the queue Q_t^B is represented with

$$\mathrm{d}Q_t^B = (a_t - c_t)\,\mathrm{d}t \quad \text{if } \begin{cases} B > Q_t^B > 0, \text{ or} \\ (Q_t^B = 0,\ a_t > c_t), \text{ or} \\ (Q_t^B = B,\ a_t < c_t), \end{cases} \tag{20}$$

and $\mathrm{d}Q_t^B = 0$, otherwise. In the following two sections we will study two important special cases of fluid queues. Our analysis is based on observing the process Q_t^B at the beginnings of the arrival process activity periods. A recent investigation of the stationary behavior of Q_t^B and its relationship to the process observed at the beginning of the activity periods can be found in [45].

4.1. Single on–off arrival process

Consider a fluid queue with capacity c and an on–off arrival process with *on* arrival rate r, $r > c$. Lengths of *on* and *off* periods are assumed to be independent i.i.d. sequences $\{\tau^{\text{on}},\ \tau_n^{\text{on}},\ n \geqslant 0\}$ and $\{\tau^{\text{off}},\ \tau_n^{\text{off}},\ n \geqslant 0\}$, respectively. Let T_n, $n \geqslant 0$, $T_0 \leqslant 0$, $T_1 > 0$ be a sequence of random times representing the beginnings of *on* periods in the arrival on–off process; $T_{n+1} - T_n = \tau_n^{\text{off}} + \tau_n^{\text{on}}$. Now, a formal construction of the on–off arrival process is as follows:

$$a_t = r \quad \text{if } t \in \big[T_n, T_n + \tau_n^{\text{on}}\big),$$

for some $n \geqslant 0$, and $a_t = 0$, otherwise. By observing the queue process Q_t^B at the beginning of *on* periods, the queue length $V_n = Q_{T_n}^B$ evolves as follows:

$$V_{n+1}^B = \big(\min\big(V_n^B + (r-c)\tau_n^{\text{on}}, B\big) - c\tau_n^{\text{off}}\big)^+, \quad n \geqslant 0. \tag{21}$$

Note that by taking $A_n = (r-c)\tau_n^{\text{on}}$ and $C_n = c\tau_n^{\text{off}}$ this recursion reduces to the recursion in (14). We assume that (21) operates in stationarity.

Next, our main object of study is the long time average loss rate for this fluid queue defined as

$$\lambda_{\text{loss}}^B \stackrel{\text{def}}{=} \lim_{t\to\infty} \frac{L(0,t)}{t}, \tag{22}$$

where $L(0,t) =$ amount of fluid lost in $(0,t)$. Here, we show that this limit exists and is equal to

$$\lambda_{\text{loss}}^B = \frac{\mathbb{E}(V_n^B + (r-c)\tau_n^{\text{on}} - B)^+}{\mathbb{E}\tau^{\text{on}} + \mathbb{E}\tau^{\text{off}}}. \tag{23}$$

Let $L_n \stackrel{\text{def}}{=} \mathbb{E}(V_n^B + (r-c)\tau_n^{\text{on}}, B)^+$, $n \geqslant 0$, be a sequence of random variables representing the losses in the renewal intervals $[T_n, T_{n+1})$, $n \geqslant 0$, respectively. If $N_t = \sup\{n\colon\ T_n < t\}$, then

$$\sum_{n=1}^{N_t-1} L_n \leqslant L(0,t) \leqslant \sum_{n=0}^{N_t} L_n. \tag{24}$$

The strong law of large numbers for renewal processes yields

$$\lim_{t\to\infty} \frac{L(0,t)}{t} = \frac{1}{\mathbb{E}\tau^{\text{on}} + \mathbb{E}\tau^{\text{off}}} \quad \text{almost surely}. \tag{25}$$

Similarly, (25) and Birkhoff's strong law of large numbers imply

$$\lim_{t\to\infty} \frac{\sum_{n=0}^{N_t} L_n}{N_t} = \mathbb{E}L_1 \quad \text{almost surely}. \tag{26}$$

Consequently, by dividing (24) with t, letting $t \to \infty$ and using (25) and (26) we derive (23).

Finally, (23) and theorem 5 yield the following asymptotic characterization of λ_{loss}^B. Let τ_e^{on} be the residual life time distribution of τ^{on}, $\mathbb{P}[\tau_e^{\text{on}} \leqslant x] = \int_0^x \mathbb{P}[\tau^{\text{on}} > u]\,\mathrm{d}u$.

Theorem 6. If $r > c$, $(r-c)\mathbb{E}\tau_{\text{on}} < c\mathbb{E}\tau_{\text{off}}$, $\tau_e^{\text{on}} \in \mathcal{S}$, then as $B \to \infty$

$$\lambda_{\text{loss}}^B = \frac{\mathbb{E}(\tau^{\text{on}}(r-c) - B)^+}{\mathbb{E}\tau^{\text{on}} + \mathbb{E}\tau^{\text{off}}}\big(1 + \text{o}(1)\big). \tag{27}$$

4.2. *Long-tailed M/G/∞ arrival process*

In this section we consider a fluid queue with capacity c and M/G/∞ arrival process. An M/G/∞ process A_t^∞ is defined by a Poisson point process with rate Λ whose points indicate the beginning of *on* periods. Each *on* period, after its activation, brings fluid with rate r to the queue for a random independent period of time τ^{on}. (For a more formal definition of an M/G/∞ process see [25].)

First, consider an indicator on–off process $\mathbf{1}(A_t^\infty > 0)$. Let I_n^{on} and I_n^{off} denote the length of the nth *on* and *off* periods, respectively. Then, it can be computed (see [25]) that

$$\mathbb{E}I_n^{\text{off}} = \frac{1}{\Lambda}, \qquad \mathbb{E}I_n^{\text{on}} = \frac{1}{\Lambda}\big(\mathrm{e}^{\Lambda\mathbb{E}\tau^{\text{on}}} - 1\big). \tag{28}$$

Furthermore, let D_n^c represent the queue increment during the nth activity period (i.e., if t_n^{b} and t_n^{e} denote the beginning and the end of the nth activity period, then $D_n^c = \int_{t_n^{\text{b}}}^{t_n^{\text{e}}} (A_t^\infty - c)\,\mathrm{d}t$). Next, under the assumption that $c \leqslant r$ the queue length V_n^B

observed at the beginning of the nth activity period of the arrival process A_t^∞ evolves according to the following recursion:

$$V_{n+1}^B = \left(\min\left(V_n^B + D_n^c, B\right) - cI_n^{\text{off}}\right)^+, \quad n \geqslant 0, \tag{29}$$

which also has the same form as the recursion in (14). Similarly, we use (22) to define the loss rate λ_{loss}^B for this fluid queue. Again, by the same arguments as in (24)–(26) we compute

$$\lambda_{\text{loss}}^B = \frac{\mathbb{E}(V_n^B + D_{n+1}^c - B)^+}{\mathbb{E}I_n^{\text{off}} + \mathbb{E}I_n^{\text{on}}}. \tag{30}$$

Now, we need to determine the asymptotic behavior of $\mathbb{P}[D_n^c > x]$ as $x \to \infty$. This behavior is known for τ^{on} being intermediately regularly varying $\tau^{\text{on}} \in \mathcal{IR}$; a non-negative random variable τ^{on} is in $\mathcal{IR} \subset \mathcal{S}$ if

$$\lim_{\eta \downarrow 1} \liminf_{x \to \infty} \frac{\mathbb{P}[\tau^{\text{on}} > \eta x]}{\mathbb{P}[\tau^{\text{on}} > x]} = 1.$$

Then, if $\tau^{\text{on}} \in \mathcal{IR}$, $0 < c < r(1 + \Lambda\mathbb{E}\tau^{\text{on}})$, [39, theorem 1] yields

$$\mathbb{P}\left[D_n^c > x\right] \sim \mathrm{e}^{\Lambda\mathbb{E}\tau^{\text{on}}}\mathbb{P}\left[\tau^{\text{on}}\eta > x\right] \quad \text{as } x \to \infty. \tag{31}$$

In the following theorem we will use the fact that $\tau^{\text{on}} \in \mathcal{IR}$, $\mathbb{E}\tau^{\text{on}} < \infty$ implies $\tau_e^{\text{on}} \in \mathcal{IR}$. Finally, the combination of (28)–(31) and theorem 5 yields the following theorem.

Theorem 7. Let $\rho = \mathbb{E}A_t^\infty = \Lambda r \mathbb{E}\tau^{\text{on}} < c$. If $c \leqslant r$ and $\tau^{\text{on}} \in \mathcal{IR}$, then

$$\lambda_{\text{loss}}^B = \Lambda\mathbb{E}\left(\tau^{\text{on}}\eta - B\right)^+\left(1 + \mathrm{o}(1)\right) \quad \text{as } B \to \infty,$$

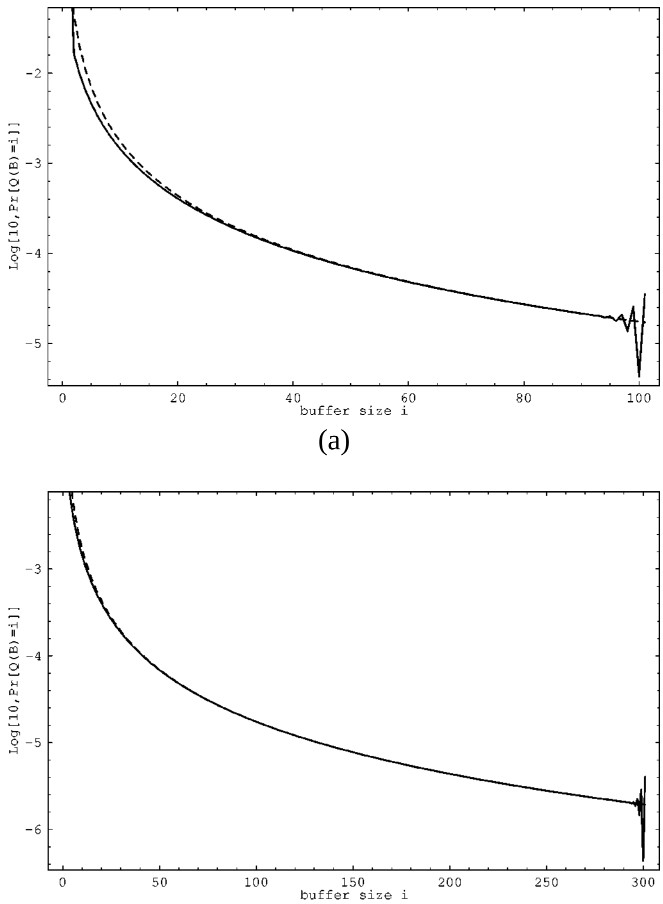
where $\eta = r + \rho - c$.

5. Numerical and simulation results

This section demonstrates, with numerical and simulation experiments, the accuracy and analytical tractability of our approximation results. The following two numerical examples will illustrate lemmas 1, 2 and theorem 4. For the case of M/G/∞ arrivals, due to the complexity of the model, we were unable to obtain a numerical solution. Thus, in examplifying theorem 7 we resort to simulation in the following subsection.

Observe that if $\mathbb{P}[A = i] \sim c/i^{\alpha+1}$ as $i \to \infty$, $\alpha > 1$, then the combination of lemmas 1 and 2 implies that for any $\varepsilon > 0$ there exists B_0 such that for all $B \geqslant B_0$

$$\left|\frac{q_i^\infty}{q_0^\infty} - \frac{q_i^B}{q_0^B}\right| \leqslant \varepsilon\frac{\delta^{B-i}}{B^{\alpha-1}}, \quad 0 \leqslant i \leqslant B, \tag{32}$$



(b)

Figure 1. Illustration for example 1.

where δ is the same as in lemma 1. The above estimate suggests that, except for i close to B, q_i^∞ is a good approximation of q_i^B. Hence, theorem 1 and (32) yield an approximation $q_i^B \approx \text{constant}/i^\alpha$ which is expected to be good for all i far enough from 0 and B. This is demonstrated in the following example.

Example 1. Consider a discrete time finite buffer queue with a constant service process $C_n \equiv 3$, and an arrival distribution $\mathbb{P}[A = 0] = 1/5$, $\mathbb{P}[A = i] = 0.6655/i^3$, $i > 0$, $\mathbb{E}A = 1.0947$. For the maximum buffer size $B = 100$ the queue occupancy probabilities are plotted with a solid line in figure 1(a). Based on (32) and theorem 1 we easily compute the suggested approximation $\tilde{q}_i^B \overset{\text{def}}{=} 0.17465/i^2$, $1 \leqslant i \leqslant B$. This approximation is plotted with dashed lines in the same figure. We can see that, with the exception of buffer sizes close to zero and $B = 100$, the approximation is very good. In fact, the relative error $|\tilde{q}_i^B - q_i^B|/q_i^B$ was smaller than 1% for the buffer

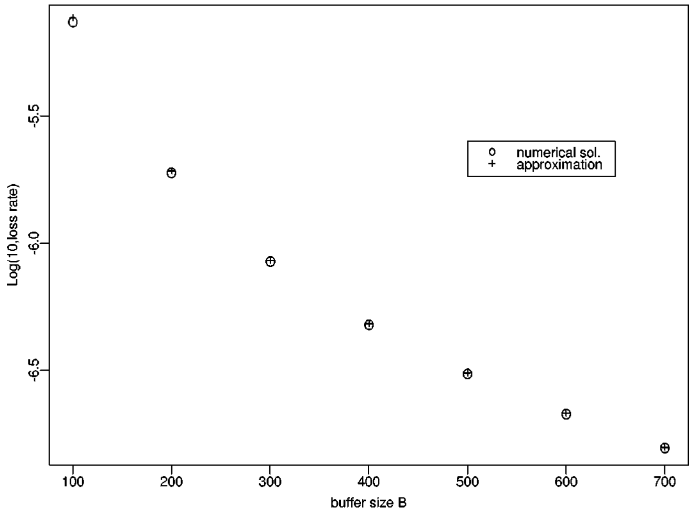


Figure 2. Illustration for example 2.

sizes $i \in [58, 93]$. Bound (32) further states that the portion of the buffer where the approximation is good increases as the maximum buffer size B increases. In addition, the length of the buffer around the boundaries where the approximation is not good stays approximately constant. To illustrate this we repeat the same experiment with the maximum buffer size $B = 300$. Now, the relative error between the actual and approximate probabilities was smaller than 1% for the buffer sizes $i \in [60, 291]$.

The next example demonstrates the accuracy of theorem 4.

Example 2. Take $C_n \equiv 2$ and an arrival distribution $\mathbb{P}[A = 0] = 1/2$, $\mathbb{P}[A = i] = 0.461969/i^4$, $i > 0$, $\mathbb{E}A = 0.5553$. Here, we numerically compute the expected loss rate $L^B_{\text{loss}} = \mathbb{E}(Q_n + A_{n+1} - 2 - B)^+$ for the maximum buffer sizes $B = 100k$, $k = 1, \dots, 7$. The results are presented with "$\circ$" symbols in figure 2. Note that for $B = 700$ we needed to solve a system of 700 linear equations. In contrast, theorem 4 readily suggests an asymptotic approximation $\tilde{L}^B_{\text{loss}} = 0.0767/B^2$. The approximation is presented in the same figure with "$+$" symbols. An excellent match is apparent from the figure. In fact, relative error $|\tilde{L}^B_{\text{loss}} - L^B_{\text{loss}}|/L^B_{\text{loss}}$ is plotted in figure 3, from which we can see that even for the smallest buffer size $B = 100$ the relative error was less than 4%.

5.1. *Fluid queue with M/G/∞ arrival process*

In this section we provide several simulation experiments to illustrate theorem 7. For simulation purposes we assume that the time is slotted with the length of a single slot being equal to one. The number of *on* periods that arrive per unit of time (slot)

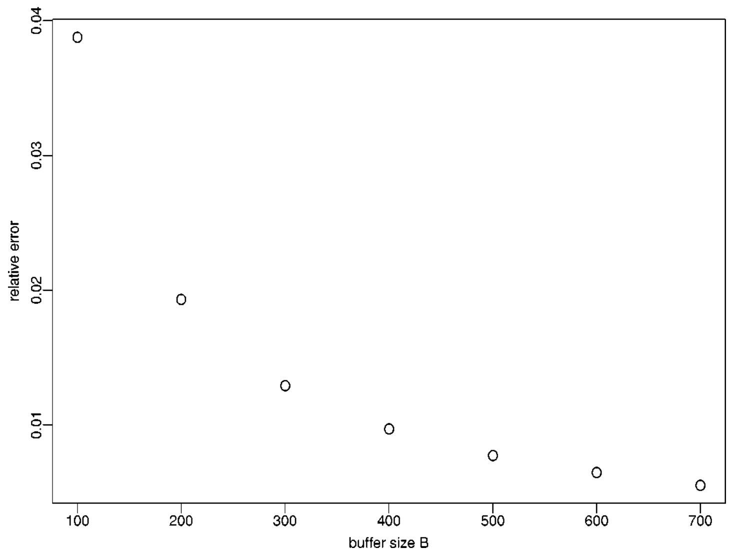


Figure 3. Illustration for example 2.

has a Poisson distribution with parameter Λ. The distribution of *on* periods is taken to be Pareto parameterized as

$$\mathbb{P}\big[\tau^{\text{on}} \geqslant n\big] = \frac{b}{b + n^{\alpha}}, \quad n = 0, 1, 2, \ldots, \ \alpha > 1, \ b > 0.$$

From this it immediately follows that the probability density behaves asymptotically as

$$\mathbb{P}\big[\tau^{\text{on}} = n\big] \sim \frac{b\alpha}{n^{\alpha+1}} \quad \text{as } n \to \infty.$$

Now, after some simple algebra we obtain

$$\tilde{\lambda}^{B}_{\text{loss}} \overset{\text{def}}{=} \Lambda \mathbb{E}\big(\tau^{\text{on}}\eta - B\big)^{+} \approx \frac{\Lambda b \eta^{\alpha}}{(\alpha - 1)B^{\alpha-1}}. \tag{33}$$

With $r = c = 1$, $b = 6$, we run two simulation experiments.

Example 3. First we choose $\Lambda = 0.35$, $\alpha = 2.5$, which implies $\mathbb{E}\tau^{\text{on}} = 2.41642$, and $\rho = 0.845747$. We simulate the losses for the maximum buffer sizes $B = 10i$, $i = 1, \ldots, 25$. The results are presented with a solid line in figure 4. In order to obtain reasonable accuracy in the experiment it was necessary to run the simulation for 2×10^9 units of time, which resulted in several days of computer processor time. Needless to say, the approximation $\tilde{\lambda}^{B}_{\text{loss}} = 0.921/B^{1.5}$ can be computed almost instantly from equation (33). The approximation $\tilde{\lambda}^{B}_{\text{loss}}$ is plotted in the same figure with dashed lines. From this figure we can see that already for the buffer size $B \approx 140$ the approximation becomes almost identical to the simulated results.

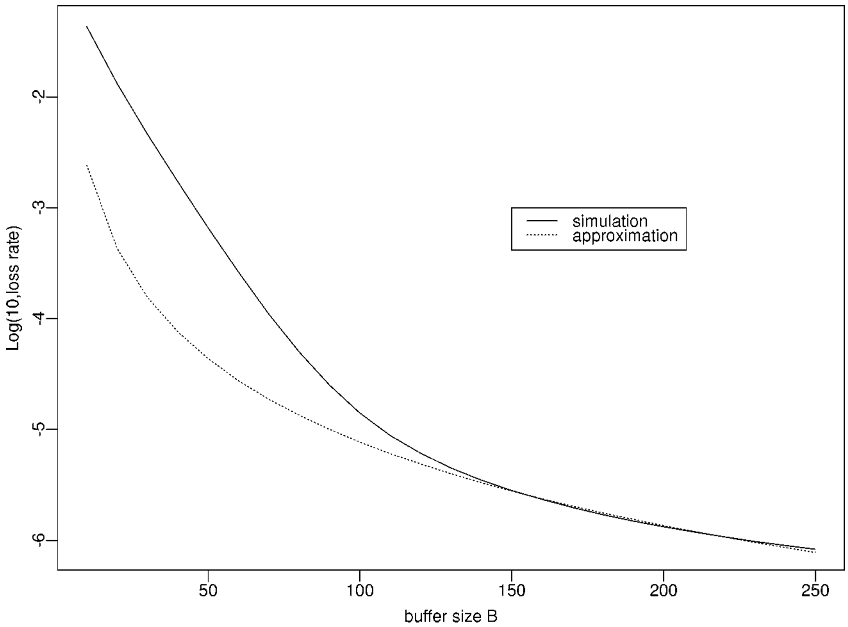


Figure 4. Illustration for example 3.

Example 4. We repeat the same experiment as the preceding one with α changed to $\alpha = 3$, which results in $\mathbb{E}\tau^{\text{on}} = 2.24304$, and $\rho = 0.785065$. The approximate loss rate computes to $\tilde{\lambda}^B_{\text{loss}} = 0.508/B^2$. An almost perfect match between the approximation and simulation results is demonstrated in figure 5. In this case the approximation becomes accurate even for smaller buffer sizes ($B \approx 80$).

6. Conclusion

In this paper we have considered several queueing systems with finite buffers and long-tailed arrivals. For these queueing systems we have derived explicit asymptotic formulas for approximating loss rates. The accuracy of the suggested approximate formulas is demonstrated on various numerical and simulation experiments. Overall, we expect that these approximate expressions, both for reasons of their explicit nature and accuracy, will be useful tools in designing modern communications switches that will be able to efficiently carry non-traditional long-tailed ("bursty") traffic.

Acknowledgements

The author is grateful to Bert Zwart for pointing out an error in the original draft of the paper and sharing the preprint of his work. We are also thankful to an anonymous reviewer for valuable suggestions.

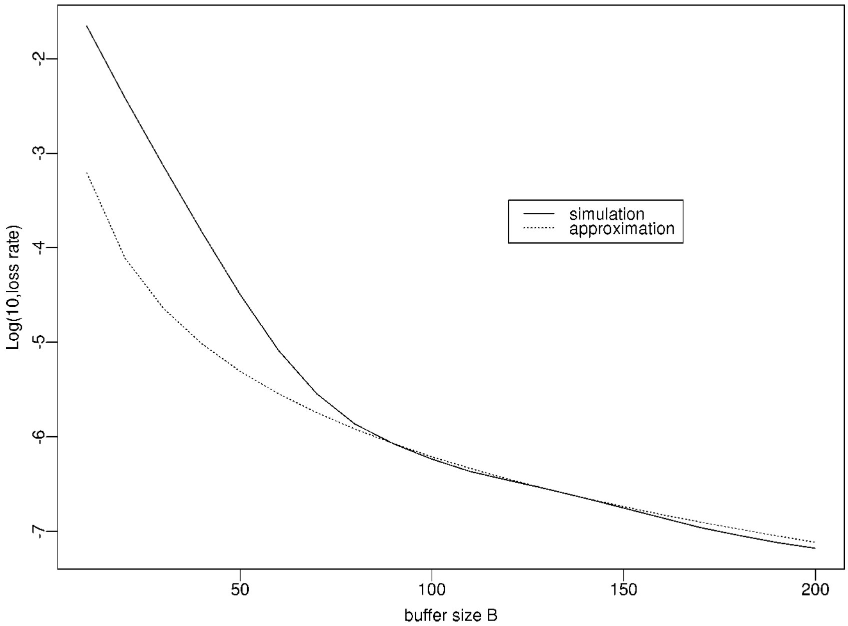


Figure 5. Illustration for example 4.

Appendix. Proofs

In this section we provide the proofs of theorem 1 and lemmas 1 and 2. Without loss of generality we assume that the set of integers that supports the distribution of $X_n = A_n - C_n$ is aperiodic and that $\mathbb{P}[X_n = -c] > 0$. Under these additional assumptions, theorem 5.2 in [3, p. 214] shows that

Claim A.1. Equation $z^c - z^c x(z) = 0$ has exactly one simple root at $z = 1$ on the unit circle $\{z\colon |z| = 1\}$ and $c - 1$ roots $z_i \neq 0$, $1 \leqslant i \leqslant c - 1$, inside the unit circle $\{z\colon |z| < 1\}$.

This fact will be repeatedly used in the following proofs.

A.1. Proof of theorem 1

First, we define subexponential probabilities (see [6, p. 429]). Let

$$p_i^{*2} \stackrel{\text{def}}{=} \sum_{k=0}^{i} p_k p_{i-k} \quad \text{and} \quad p_i^{*n} \stackrel{\text{def}}{=} \sum_{k=0}^{i} p_k^{*(n-1)} p_{i-k}$$

denote two-fold and n-fold convolution of p_i, respectively.

Definition A.1. A non-negative sequence $\{p_i,\ i \geqslant 0\}$ is called *long-tailed* if for any integer k,

$$\lim_{i\to\infty} \frac{p_{i+k}}{p_i} = 1.$$

If, in addition, $\{p_i,\ i \geqslant 0\}$ is a probability sequence ($\sum_{i=0}^{\infty} p_i = 1$), and

$$\lim_{i\to\infty} \frac{p_i^{*2}}{p_i} = 2,$$

we say that p_i is *subexponential density* $p_i \in \mathcal{S}^d$.

It is easy to check that for an integer valued random variable A for which $A \in \mathcal{S}^*$, the probability sequence $p_i = \mathbb{P}[A \geqslant i]/\mathbb{E}A$, $i \geqslant 1$, belongs to $\mathcal{S}^d$.

By using a well-known connection between the queue length distribution and the supremum of the corresponding random walk with increments $X_n = A_n - C_n$, the pgf $q(z)$ of Q_n can be represented as (see [17, chapter XII])

$$q(z) = \frac{1 - g_+(1)}{1 - g_+(z)}, \tag{A.1}$$

where $g_+(z) = \sum_{i=1}^{\infty} g_{+i} z^i$ is the generating function of a strictly ascending ladder height random variable for which $g_+(1) < 1$ iff $\mathbb{E}X_n < 0$. Equation (A.1) can be written in its equivalent form

$$q_i = \bigl(1 - g_+(1)\bigr) \sum_{k=0}^{\infty} g_{+i}^{*k}, \tag{A.2}$$

where g_{+i}^{*0} is a unit mass at zero. In a subexponential framework, the asymptotic behavior of the random sum (A.2) is characterized as follows.

Lemma A.1. If $g_+(1) < 1$, and $g_{+i}/g_+(1) \in \mathcal{S}^d$ then

$$\lim_{i\to\infty} \frac{q_i}{g_{+i}} = \frac{1}{1 - g_+(1)}.$$

Remark. An equivalent result for random sums of continuous subexponential densities can be found in [28].

Proof. Follows from [11, lemma 5], [16, lemma 2], and dominated convergence. □

At this point, it is clear that in order to establish the asymptotic connection between the arrival distribution and the queue length distribution we need to investigate the asymptotic behavior of g_{+i}. This is presented in the following lemma.

Lemma A.2. If $\mathbb{P}[A > i]$ is long-tailed, $\mathbb{P}[C \leqslant c] = 1$, $c < \infty$, and $\mathbb{E}A < \mathbb{E}C$, then

$$g_{+i} \sim \frac{1 - g_{+}(1)}{\mathbb{E}C - \mathbb{E}A}\mathbb{P}[A > i] \quad \text{as } i \to \infty.$$

Remark. In [35,43], the asymptotic behavior for $\sum_{k=i}^{\infty} g_{+k}$ was obtained; their proving technique uses directly the monotonicity of $\sum_{k=i}^{\infty} g_{+k}$. Since g_{+i} is not necessarily monotonic, we were unable to adopt this method of proof here.

Proof. Standard derivation in queueing theory (e.g., see [33]) shows that the pgf $q(z)$ of Q_n satisfies

$$q(z) = \frac{\sum_{i=0}^{c-1} \sum_{k=0}^{i} q_k x_{i-k-c}(z^c - z^i)}{z^c - z^c x(z)}, \tag{A.3}$$

where $x_i = \mathbb{P}[A_n - C_n = i]$, and $x(z) = \sum_{i=-c}^{\infty} x_i z^i$. Then by claim A.1, equation $z^c - z^c x(z) = 0$ has a simple root at $z = 1$ and $c - 1$ roots $z_i, 1 \leqslant i \leqslant c - 1$, inside the unit circle. Since $q(z)$ is an analytic function in the unit circle these roots must be zeros of the numerator in (A.3), i.e.,

$$\sum_{i=0}^{c-1} \sum_{k=0}^{i} q_k x_{i-k-c}\left(z^c - z^i\right) = h(z-1)(z-z_1)\cdots(z-z_{c-1}), \tag{A.4}$$

for some constant h. Next, using (A.4) in evaluating the right hand side of (A.3) (by l'Hospital's rule) and equating it to $q(1) = 1$ we derive

$$h(1-z_1)\cdots(1-z_{c-1}) = \mathbb{E}C - \mathbb{E}A. \tag{A.5}$$

Further, by equating (A.1) and (A.3) we obtain

$$g_+(z) = \frac{(1 - g_+(1))(z^c x(z) - z^c)}{h(z-1)(z-z_1)\cdots(z-z_{c-1})} + 1. \tag{A.6}$$

Now, examine the coefficients of the analytic function

$$x^1(z) \stackrel{\text{def}}{=} \frac{z^c x(z) - z^c}{z - 1} = \frac{z^c x(z) - 1}{z - 1} - \sum_{i=0}^{c-1} z^i.$$

The coefficients x_i^1 of the analytic expansion of $x^1(z)$, $|z| \leqslant 1$, for $i > c - 1$, are positive and satisfy

$$x_i^1 = \sum_{k=i+1}^{\infty} x_{i-c} = \sum_{j=0}^{c} c_j \mathbb{P}[A > i + j - c],$$

from which, by the assumption that $\mathbb{P}[A > i]$ is long-tailed it easily follows that

$$x_i^1 \sim \mathbb{P}[A > i] \quad \text{as } i \to \infty. \tag{A.7}$$

Next, we investigate the coefficients x_i^2 of the analytic expansion of

$$x^2(z) \stackrel{\text{def}}{=} \frac{x^1(z)}{z - z_1},$$

where z_1 is already defined to be one of the roots from the unit circle. By expanding $1/(z - z_1)$ into an analytic series in the strip $\delta < |z| < 1$, $\delta \stackrel{\text{def}}{=} \max |z_i| < 1$, we arrive at

$$x^2(z) = x^1(z) \frac{1}{z} \sum_{i=0}^{\infty} \left(\frac{z_1}{z} \right)^i,$$

which, after comparing the coefficients on the left- and right-hand side, yields

$$x_i^2 = \sum_{k=i+1}^{\infty} x_k^1 z_1^{k-i-1}.$$

Here, for any constant $K \in \mathbb{N}$, $K > c$,

$$\begin{aligned} \left| \frac{x_i^2}{x_i^1} - \frac{1}{1 - z_1} \right| &\leqslant \sum_{k=0}^{K} \left| \frac{x_{i+k+1}^1}{x_i^1} - 1 \right| |z_1|^k + \sum_{k=K+1}^{\infty} \left| \frac{x_{i+k+1}^1}{x_i^1} - 1 \right| |z_1|^k \\ &\leqslant \sum_{k=0}^{K} \left| \frac{x_{i+k+1}^1}{x_i^1} - 1 \right| |z_1|^k + \frac{2|z_1|^K}{1 - |z_1|}, \end{aligned} \tag{A.8}$$

where in the last inequality we have used the monotonicity of x_i^1 ($i > c$), i.e., $|x_{i+k+1}^1 / x_i^1 - 1| \leqslant x_{i+k+1}^1 / x_i^1 + 1 \leqslant 2$. Inequality (A.8) readily implies

$$\lim_{i \to \infty} \frac{x_i^2}{x_i^1} = \frac{1}{1 - z_1}.$$

Similarly, by repeating this procedure for the remaining $c - 2$ roots, we derive the asymptotic behavior of the coefficients of $x^{j+1}(z) \stackrel{\text{def}}{=} x^j(z)/(z - z_j)$. Note that for $j \geqslant 2$, a similar bound as in (A.8) can be obtained using $|x_i^j| \leqslant x_i^1/(1 - \delta)^j$, and the monotonicity of x_i^1 for $i > c$. Thus, we have derived that

$$x_i^c \sim \frac{\mathbb{P}[A > i]}{(1 - z_1) \cdots (1 - z_{c-1})} \quad \text{as } i \to \infty. \tag{A.9}$$

By combining (A.9) with (A.5), and by equating coefficients on the left- and right-hand side in (A.6), we obtain the conclusion of the lemma. □

Finally, the proof of theorem 1 follows from lemmas A.1 and A.2.

A.2. Proof of lemma 1

The case $c > 1$ is quite involved due to the more complicated boundary conditions. Define

$$s^B(z) \stackrel{\text{def}}{=} \sum_{i=0}^{c-1} \sum_{k=0}^{i} \nu_k^B x_{i-k-c}\big(z^c - z^i\big), \quad B \in \mathbb{N} \cup \{\infty\}.$$

Since $q^B(z)$, $B \leqslant \infty$, is an analytic function for $|z| \leqslant 1$, it follows that the numerator and denominator of (6) have exactly the same zeros. Recall from claim A.1 that $z^c - z^c x(z) = 0$ has a simple zero at $z = 1$ and $c - 1$ zeros inside the unit circle ($z_i \neq 0$). Assume that l, $l \leqslant c - 1$, of the zeros from within the unit circle are distinct with multiplicities m_i, $1 \leqslant i \leqslant l$, $\sum_{i=1}^{l} m_i = c - 1$. Then, ν_k^B, $1 \leqslant k \leqslant c - 1$, satisfy the following set of $c - 1$ equations:

$$s^{B(n)}(z_k) + \frac{R^{B(n)}(z_k)}{q_0^B} = 0, \quad 1 \leqslant k \leqslant l,\ 0 \leqslant n \leqslant m_k - 1, \tag{A.10}$$

where $s^{B(n)}(z) = \mathrm{d}^n s^B(z)/\mathrm{d}z^n$ and $R^{B(n)}(z) = \mathrm{d}^n R^B(z)/\mathrm{d}z^n$. Similarly, ν_k^∞, $1 \leqslant k \leqslant c - 1$, satisfy

$$s^{\infty(n)}(z_k) = 0, \quad 1 \leqslant k \leqslant l,\ 0 \leqslant n \leqslant m_k - 1, \tag{A.11}$$

with $s^{\infty(n)}(z) = \mathrm{d}^n s^\infty(z)/\mathrm{d}z^n$. By subtracting (A.10) from (A.11) we obtain that $(\nu_k^\infty - \nu_k^B)$, $1 \leqslant k \leqslant c - 1$, is a solution to

$$s^{\infty(n)}(z_k) - s^{B(n)}(z_k) = \frac{R^{B(n)}(z_k)}{q_0^B}, \quad 1 \leqslant k \leqslant l,\ 0 \leqslant n \leqslant m_k - 1. \tag{A.12}$$

Next, let $\Delta\nu$ and R be two column vectors with corresponding elements $\nu_k^\infty - \nu_k^B$, $1 \leqslant i \leqslant c - 1$, and $R^{B(n)}(z_k)/q_0^B$, $1 \leqslant k \leqslant l$, $0 \leqslant n \leqslant m_k - 1$, and let X and Z be matrices with rows equal to

$$(x_{i-1-c}, \ldots, x_{-c}, 0, \ldots, 0), \quad 1 \leqslant i \leqslant c - 1,$$

and

$$\left(\frac{\mathrm{d}}{\mathrm{d}z_k}\right)^n \big(z_k^c - z_k, \ldots, z_k^c - z_k^{c-1}\big), \quad 1 \leqslant k \leqslant l,\ 0 \leqslant n \leqslant m_k - 1,$$

respectively. Then, (A.12) can be expressed in the following compact form:

$$ZX\Delta\nu = R. \tag{A.13}$$

Here, observe that $\det(X) = (x_c)^{c-1} > 0$; also, by using the basic properties of determinants, for the case when all of the roots are distinct ($l = c - 1$), we compute

$$\det(Z) = \prod_{i=1}^{c-1} z_i(z_i - 1) \prod_{c-1 \geqslant j > k \geqslant 1} (z_j - z_k) \neq 0. \tag{A.14}$$

When some of the roots have multiplicities bigger than one (i.e., $l < c-1$), the determinant of Z can be computed by taking the corresponding derivatives in (A.14) and then equating the appropriate roots. This elementary computation, the details of which have been omitted, yields

$$\left|\det(Z)\right| = N_c \prod_{i=1}^{l} \left|z_i(z_i-1)\right|^{m_i} \prod_{l\geqslant j>k\geqslant 1} \left|z_i(z_i-1)\right|^{m_j m_k} > 0,$$

where $N_c = \prod_{i=1}^{l}\prod_{j=0}^{m_i-1} j!$. Therefore, X and Z are nonsingular matrices and from (A.13) we derive

$$\Delta\nu = X^{-1}Z^{-1}R. \tag{A.15}$$

Furthermore, $R^B(z_k)$, as defined in (7), and its derivatives satisfy

$$\begin{aligned} \left|R^{B(n)}(z_k)\right| &= \left|\left(\frac{\mathrm{d}}{\mathrm{d}z_k}\right)^n \sum_{i=B+1}^{\infty}\sum_{j=0}^{B} q_j^B x_{i-j}\left(z_k^{B+c} - z_k^{i+c}\right)\right| \\ &= \mathrm{O}\left(|z_k|^B \mathbb{P}\left[Q_n^B + A_{n+1} - C_{n+1} > B\right]\right), \quad n \geqslant 0. \end{aligned} \tag{A.16}$$

Also, the stochastic dominance $\mathbb{P}[Q_n^\infty > 0] \geqslant \mathbb{P}[Q_n^B > 0]$ implies

$$q_0^B \geqslant q_0^\infty > 0. \tag{A.17}$$

Therefore, by replacing (A.16) and (A.17) in (A.15) we arrive at

$$\max_{1\leqslant k\leqslant c-1}\left|\nu_k^B - \nu_k^\infty\right| = \mathrm{O}\left(\delta^B \mathbb{P}\left[Q_n^B + A_{n+1} - C_{n+1} > B\right]\right), \tag{A.18}$$

where $\delta = \max_{1\leqslant k\leqslant c-1}|z_k|$.

Next, from (6), $q^B(z)$ can be rewritten as

$$\frac{q^B(z)}{q_0^B} - \frac{q^\infty(z)}{q_0^\infty} = \frac{q^\infty(z)}{q_0^\infty}\left[\frac{\Delta s^B(z)}{s^\infty(z)} + \frac{R^B(z)}{q_0^B s^\infty(z)}\right], \tag{A.19}$$

where $\Delta s^B(z) \stackrel{\text{def}}{=} s^B(z) - s^\infty(z)$. Note that the first $B+1$ coefficients in the analytic expansion of the expression on the left-hand side of equation (A.19) are equal to $\nu_k^B - \nu_k^\infty$, $0 \leqslant k \leqslant B$. Thus, an estimate of the coefficients in the analytic expansion of the expression on the right-hand side of (A.19) will yield a bound on $|\nu_k^B - \nu_k^\infty|$. This estimate will be obtained in the remainder of the proof.

First, observe that $R^B(z)$ has a zero of multiplicity $B+c$ at $z=0$ and $s^\infty(z)$ is a polynomial of degree c with no zeros at $z=0$. Consequently, $R^B(z)/s^\infty(z)$ is analytic in some neighborhood of $z=0$ and it has a zero of order at least $B+c$ at $z=0$. Therefore, its analytic expansion $R(z)/s^\infty(z) = \sum_{i=0}^\infty b_{1i}z^i$ at $z=0$ has its first $B+c+1$ coefficients equal to zero, i.e.,

$$b_{1i} = 0, \quad 0 \leqslant i \leqslant B+c. \tag{A.20}$$

Similarly, $\Delta s^B(z)/s^\infty(z)$ is analytic in some neighborhood of $z = 0$ with its analytic expansion denoted as $\sum_{i=0}^{\infty} b_{2i} z^i$. Now, we intend to bound the coefficients b_{2i}, $0 \leqslant i \leqslant B$. From (A.18) it easily follows that

$$b_{2i} = \frac{\mathrm{d}^i}{\mathrm{d}z^i} \frac{\Delta s^B(z)}{s^\infty(z)}\bigg|_{z=0} = \mathrm{O}\big(\delta^B \mathbb{P}\big[Q_n^B + A_{n+1} - C_{n+1} > B\big]\big), \quad 0 \leqslant i \leqslant c-1. \tag{A.21}$$

Next, notice that $\Delta s^B(1) = s^\infty(1) = 0$, and thus, $\Delta s_1^B(z) \stackrel{\text{def}}{=} s^B(z)/(z-1)$ is a polynomial of $c-1$ degree and

$$\frac{z-1}{s^\infty(z)} = \sum_{k=1}^{l} \sum_{n=1}^{m_k} \frac{D_{kn}}{(z-z_k)^n},$$

for some complex constants D_{kn}. Using this, we arrive at

$$\frac{\Delta s^B(z)}{s^\infty(z)} = \sum_{k=1}^{l} \sum_{n=1}^{m_k} D_{kn} \frac{\Delta s_1^B(z)}{(z-z_k)^n}. \tag{A.22}$$

Then, $\Delta s_1^B(z)/(z-z_k)^n$ is analytic for $|z| < |z_k|$ and it can be expanded at $z = 0$ into an analytic series

$$\frac{\Delta s_1^B(z)}{(z-z_k)^n} = \sum_{i=0}^{\infty} b_{2kn(i)} z^i, \quad 1 \leqslant k \leqslant l,\ 1 \leqslant n \leqslant m_k. \tag{A.23}$$

By replacing the identity

$$\begin{aligned}\frac{1}{(z-z_k)^n} &= \frac{1}{(n-1)!}\left(\frac{\mathrm{d}}{\mathrm{d}z_k}\right)^{n-1} \frac{1}{z-z_k} \\ &= -\frac{1}{(n-1)!}\left(\frac{\mathrm{d}}{\mathrm{d}z_k}\right)^{n} \frac{1}{z_k} \sum_{j=0}^{\infty} (z_k)^{-j} z^j, \quad |z| < |z_k|,\end{aligned}$$

in (A.23) and then equating the coefficients next to z^i, $i \geqslant c$, on both sides of (A.23) we compute

$$b_{2kn(i)} = -\frac{1}{(n-1)!}\left(\frac{\mathrm{d}}{\mathrm{d}z_k}\right)^{n} \frac{s_1^B(z_k)}{(z_k)^{i+1}} = \frac{1}{(n-1)!}\left(\frac{\mathrm{d}}{\mathrm{d}z_k}\right)^{n} \frac{R^B(z_k)}{(z_k)^{i+1}}.$$

By combining the preceding equation with (A.16) we conclude that for $i \geqslant c$

$$\big|b_{2kn(i)}\big| = \mathrm{O}\big(\delta^{B-i}\, \mathbb{P}\big[Q_n^B + A_{n+1} - C_{n+1} > B\big]\big). \tag{A.24}$$

Thus, (A.21), (A.22) and (A.24) yield

$$|b_{2i}| = \mathrm{O}\big(\delta^{B-i}\, \mathbb{P}\big[Q_n^B + A_{n+1} - C_{n+1} > B\big]\big), \quad 0 \leqslant i \leqslant B. \tag{A.25}$$

Finally, from (A.19), (A.20) and (A.25) we compute

$$\begin{aligned}
\left|\nu_k^B - \nu_k^\infty\right| &= \left|\frac{1}{q_0^\infty}\sum_{i=0}^{k} b_{2i} q_{k-i}^\infty\right| \\
&\leqslant \frac{1}{q_0^\infty}\sum_{i=0}^{k} |b_{2i}| \\
&= \frac{1}{q_0^\infty}\sum_{i=0}^{k} \mathrm{O}\left(\delta^{B-i}\,\mathbb{P}\left[Q_n^B + A_{n+1} - C_{n+1} > B\right]\right) \\
&= \mathrm{O}\left(\delta^{B-k}\mathbb{P}\left[Q_n^B + A_{n+1} - C_{n+1} > B\right]\right).
\end{aligned}$$

This finishes the proof of the lemma. $\square$

A.3. *Proof of lemma 2*

The idea for proving this lemma is to stochastically bound the finite buffer queue length random variable Q_n^B with an infinite buffer queue having truncated arrivals. More precisely, let us define a sequence of truncated random variables

$$A_n^B = \min(A_n, B + c), \quad n \geqslant 0.$$

Let $Q_n^{B,B}$ be a queueing process characterized with a finite buffer B and truncated arrival process A_n^B. Then, by assuming that $Q_0^{B,B} = Q_0^B$, and by using induction in n,

$$\begin{aligned}
Q_{n+1}^B &= \min\left(\left(Q_n^B + A_{n+1} - C_{n+1}\right)^+, B\right) \\
&= \min\left(\left(Q_n^B + A_{n+1}^B - C_{n+1}\right)^+, B\right) = Q_{n+1}^{B,B},
\end{aligned} \tag{A.26}$$

it follows that $Q_n^{B,B} = Q_n^B$, $n \geqslant 0$. Furthermore, if we denote by $Q_n^{\infty,B}$, $n \geqslant 0$, $Q_0^{\infty,B} = Q_0^B$, a queueing process of an infinite buffer queue with truncated arrival sequence A_n^B

$$Q_{n+1}^{\infty,B} = \left(Q_n^{\infty,B} + A_{n+1}^B - C_{n+1}\right)^+,$$

then, again by induction in n, it follows that

$$Q_n^B = Q_n^{B,B} \leqslant Q_n^{\infty,B}, \quad n \geqslant 0.$$

Thus,

$$\begin{aligned}
\mathbb{P}\left[Q_n^B + A_{n+1} - C_{n+1} \geqslant B\right] &= \mathbb{P}\left[Q_n^B + A_{n+1}^B - C_{n+1} \geqslant B\right] \\
&\leqslant \mathbb{P}\left[Q_n^{\infty,B} + A_{n+1}^B - C_{n+1} \geqslant B\right] \\
&= \mathbb{P}\left[Q_{n+1}^{\infty,B} \geqslant B\right].
\end{aligned} \tag{A.27}$$

Therefore, by passing $n \to \infty$ in (A.27), we obtain that in stationarity

$$\mathbb{P}\left[Q_n^B + A_{n+1} - C_{n+1} \geqslant B\right] \leqslant \mathbb{P}\left[Q_n^{\infty,B} \geqslant B\right].$$

Hence, the proof of lemma 2 will follow if we show that the stationary distribution of $Q_n^{\infty,B}$ satisfies

$$\mathbb{P}\big[Q_n^{\infty,B} > B\big] = \mathrm{o}\big(\mathbb{P}[A_e > B]\big) \quad \text{as } B \to \infty. \tag{A.28}$$

Now, in order to prove (A.28) we first need to derive preliminary lemmas A.3 and A.5.

Similarly as in equation (A.3), the pgf $q^{\infty,B}(z)$ of $Q_n^{\infty,B}$ satisfies

$$q^{\infty,B}(z) = \frac{\sum_{i=0}^{c-1}\sum_{k=0}^{i} q_k^{\infty,B} x_{i-k-c}^B (z^c - z^i)}{z^c - z^c x^B(z)}, \tag{A.29}$$

where $x_i^B = \mathbb{P}[A_n^B - C_n = i]$ and $x^B(z) = \sum_{i=-c}^{B+c} x_i^B z^i$. The above expression can be written in its equivalent form (the same as (A.1))

$$q^{\infty,B}(z) = \frac{1 - g_+^B(1)}{1 - g_+^B(z)}, \tag{A.30}$$

where $g_+^B(z) = \sum_{i=1}^{B+c} g_{+i}^B z^i$ is the pgf of a strictly ascending ladder height random variable.

Lemma A.3. If $\mathbb{E}A < \mathbb{E}C$, then there exists a universal constant $K_3 > 0$ such that for all $B > 0$,

$$g_{+i}^B \leqslant K_3 \mathbb{P}[A_n \geqslant i - c], \quad 1 \leqslant i \leqslant B + c.$$

Proof. Again, claim A.1 shows that $z^c - z^c x^B(z) = 0$ has a simple zero at $z = 1$ and $c - 1$ zeros $z_i(B)$, $1 \leqslant i \leqslant c - 1$, strictly inside the unit circle. These zeros are also the zeros of the polynomial in the numerator of (A.29) which we denote as

$$N^B(z) \overset{\text{def}}{=} \sum_{i=0}^{c-1}\sum_{k=0}^{i} q_k^{\infty,B} x_{i-k-c}\big(z^c - z^i\big);$$

note that

$$x_i^B = \mathbb{P}\big[\min(A, B + c) - C = i\big] = \mathbb{P}[A - C = i] \equiv x_i, \quad -c \leqslant i < B.$$

Furthermore, since the d.f. of A^B converges to the d.f. of A as $B \to \infty$, theorem 1 of [7, p. 219] implies that for any fixed k

$$q_k^{\infty,B} \to q_k \quad \text{as } B \to \infty, \tag{A.31}$$

where q_k, $k \geqslant 0$, are the queue stationary probabilities that correspond to the (non-truncated) arrival sequence A_n. Therefore,

$$N^B(z) \to N(z) \overset{\text{def}}{=} \sum_{i=0}^{c-1}\sum_{k=0}^{i} q_k x_{i-k-c}\big(z^c - z^i\big),$$

as $B \to \infty$. Hence, the zeros of $N^B(z)$ from within the unit circle converge to the corresponding zeros z_i, $1 \leqslant i \leqslant c-1$, of $N(z)$ which are also inside the unit circle. If $\delta = \max_{1 \leqslant i \leqslant c-1} |z_i| < 1$, then for any $0 < \varepsilon < \varepsilon + \delta < 1$ we can choose B_0 large enough such that for all $B \geqslant B_0$ all of the zeros $z_i(B)$, $1 \leqslant i \leqslant c-1$, are in the ε neighborhood of one of the zeros z_i, $1 \leqslant i \leqslant c-1$. This implies that

$$\sup_{B \geqslant B_0} \max_{1 \leqslant i \leqslant c-1} \bigl|z_i(B)\bigr| < \delta + \varepsilon < 1,$$

and, since all of the zeros for $B < B_0$ are strictly within the unit circle, we conclude that there exists a constant $0 < \delta_1 < 1$, such that

$$\sup_{B \geqslant 0} \max_{1 \leqslant i \leqslant c-1} \bigl|z_i(B)\bigr| \leqslant \delta_1 < 1.$$

Next, observe that by equating (A.30) with (A.29) we compute

$$g_+^B(z) = \frac{(1 - q_+^B(1))(z^c x^B(z) - z^c)}{h^B(z-1)(z - z_1(B)) \cdots (z - z_{c-1}(B))} + 1, \tag{A.32}$$

where

$$h^B(z-1)\bigl(z - z_1(B)\bigr) \cdots \bigl(z - z_{c-1}(B)\bigr) = N^B(z),$$

and the constant

$$h^B = \sum_{i=0}^{c-1} \sum_{k=0}^{i} q_k^{\infty,B} x_{i-k-c}.$$

Hence, we can obtain $g_+^B(z)$ by canceling the zeros in the numerator and denominator of (A.32). First, let us examine the coefficients of the polynomial

$$x^{B,1}(z) \stackrel{\text{def}}{=} \frac{z^c x^B(z) - z^c}{z-1} = \frac{z^c x^B(z) - 1}{z-1} - \sum_{i=0}^{c-1} z^i.$$

The coefficients of this polynomial $x_i^{B,1}$, for $c - 1 < i \leqslant B + 2c - 1$, are positive and satisfy

$$x_i^{B,1} = \sum_{k=i+1}^{B+2c} x_{k-c}^B \leqslant \sum_{j=0}^{c} \mathbb{P}[C = j]\, \mathbb{P}[A > i + j - c] \leqslant \mathbb{P}[A > i - c].$$

Next, we investigate the coefficients of

$$x^{B,2}(z) \stackrel{\text{def}}{=} \frac{x^{B,1}(z)}{z - z_1(B)},$$

where $z_1(B)$ is already defined to be one of the roots from inside the unit circle. By expanding $1/(z - z_1(B))$ into an analytic series in the strip $\delta_1 < |z| < 1$ we arrive at

$$z^{B,2}(z) = x^{B,1}(z)\,\frac{1}{z}\sum_{i=0}^{\infty}\left(\frac{z_1(B)}{z}\right)^i,$$

which, after comparing the coefficients on the left- and right-hand side, yields

$$x_i^{B,2} = \sum_{k=i+1}^{B+2c-1} x_k^{B,1}\big(z_1(B)\big)^{k-i-1}.$$

This equality readily implies

$$\left|x_i^{B,2}\right| \leqslant \sum_{k=i+1}^{\infty}\left|x_k^{B,1}\right|\delta_i^{k-i-1} \leqslant \sum_{k=i+1}^{\infty}\mathbb{P}[A \geqslant k-c]\delta_1^{k-i-1}, \quad i \geqslant c.$$

Finally, the monotonicity of $\mathbb{P}[A \geqslant i]$ yields

$$\left|x_i^{B,2}\right| \leqslant \frac{1}{1-\delta_1}\mathbb{P}[A \geqslant i-c], \quad i \geqslant c.$$

Consequently, by repeating this procedure for the remaining $c-2$ roots we derive

$$\left|x_i^{B,c}\right| \leqslant \frac{1}{(1-\delta_1)^{c-1}}\mathbb{P}[A \geqslant i-c], \quad i \geqslant c.$$

Finally, by combining the preceding bound and (A.32) we derive

$$g_{+i}^B \leqslant \frac{1}{h^B}\,\frac{1}{(1-\delta_1)^{c-1}}\mathbb{P}[A \geqslant i-c], \quad i \geqslant c. \tag{A.33}$$

Furthermore, by (A.31)

$$\lim_{B\to\infty} h^B = \sum_{i=0}^{c-1}\sum_{k=0}^{i} q_k x_{i-k-c} \geqslant q_0 x_{-c} > 0,$$

and therefore $1/h^B = \mathrm{O}(1)$, which when replaced in (A.33) completes the proof of lemma A.3. □

At this point, in order to prove lemma A.5, we will need the following estimate on the distribution of a sum with uniformly bounded summands.

Lemma A.4. Let X_i, $i \geqslant 1$, be a sequence of non-negative i.i.d. random variables and let $S_n = \sum_{i=1}^n X_i$, $n \geqslant 1$. If $X_1 \in \mathcal{S}$ then for any fixed n, $c < \infty$

$$\mathbb{P}\Big[S_n \geqslant B,\ \max_{1\leqslant i\leqslant n} X_i \leqslant B+c\Big] = \mathrm{o}\big(\mathbb{P}[X_1 \geqslant B]\big) \quad \text{as } B \to \infty.$$

Proof. Observe that $\mathbb{P}[S_n \geqslant B]$ can be decomposed as follows:

$$\begin{aligned}\mathbb{P}[S_n \geqslant B] &= n\mathbb{P}\Big[S_n \geqslant B, X_n > B+c, \max_{1\leqslant i\leqslant n-1} X_i \leqslant B+c\Big]\\ &\quad + \mathbb{P}\Big[S_n \geqslant B, \bigcup_{1\leqslant i<k\leqslant n}\{X_i > B+c, X_k > B+c\}\Big]\\ &\quad + \mathbb{P}\Big[S_n \geqslant B, \max_{1\leqslant i\leqslant n} X_i \leqslant B+c\Big]\\ &\stackrel{\text{def}}{=} nP_{1n} + P_{2n} + P_{3n}.\end{aligned} \tag{A.34}$$

From the definition of subexponential distributions it follows that

$$\mathbb{P}[S_n \geqslant B] \sim n\mathbb{P}[X_1 \geqslant B] \quad \text{as } B \to \infty. \tag{A.35}$$

Next, note that

$$\begin{aligned}P_{1n} &= \mathbb{P}\Big[X_n > B+c, \max_{1\leqslant i\leqslant n-1} X_i \leqslant B+c\Big] = \mathbb{P}[X_n > B+c]\mathbb{P}[X_n \leqslant B+c]^{n-1}\\ &\sim \mathbb{P}[X_1 \geqslant B] \quad \text{as } B \to \infty.\end{aligned} \tag{A.36}$$

Also,

$$P_{2n} \leqslant \frac{n(n-1)}{2}\mathbb{P}[X_n > B]^2 = \mathrm{o}\big(\mathbb{P}[X_n > B]\big) \quad \text{as } B \to \infty. \tag{A.37}$$

Finally, by replacing (A.35)–(A.37) in (A.34) we obtain

$$P_{3n} = \mathrm{o}\big(\mathbb{P}[X_n > B]\big) \quad \text{as } B \to \infty,$$

which concludes the proof of the lemma. □

Next, let $\{X^B,\ X_i^B,\ i \geqslant 1\}$ be a family of integer valued positive i.i.d. random variables with probabilities $g_i^B = \mathbb{P}[X^B = i]$. For each B, X^B has a finite support $\mathbb{P}[X^B \leqslant B+c] = 1$, for some fixed constant c.

Lemma A.5. Let $S_n^B = \sum_{i=1}^n X_i^B$, $n \geqslant 1$. If for all B, $g_i^B \leqslant K\mathbb{P}[A \geqslant i-c]$, $1 \leqslant i \leqslant B+c$, $K > 0$, and $A_e \in \mathcal{S}$ then

(i) for any fixed n,

$$\mathbb{P}\big[S_n^B \geqslant B\big] = \mathrm{o}\big(\mathbb{P}[A_e \geqslant B]\big) \quad \text{as } B \to \infty.$$

(ii) for any $\varepsilon > 0$ there exist a finite constant K_ε such that for all B and $n \geqslant 1$,

$$\mathbb{P}\big[S_n^B \geqslant B\big] \leqslant K_\varepsilon(1+\varepsilon)^n\mathbb{P}[A_e \geqslant B].$$

Proof. Let X_i, $i \geqslant 1$, be sequence of non-negative, integer valued random variables with probability mass function

$$\mathbb{P}[X_k = i] = \mathbb{P}[A \geqslant i - c]/\mathbb{E}A_c, \quad i \geqslant 1, \ \mathbb{E}A_c = \sum_{i=1}^{\infty} \mathbb{P}[A \geqslant i - c]$$

and let $S_n = \sum_{i=1}^{n} X_i$ be their partial sums. Then, $X_i \in S$ and it easily follows that

$$\mathbb{P}\big[S_n^B \geqslant B\big] \leqslant (K\mathbb{E}A)^n \mathbb{P}\Big[S_n \geqslant B, \ \max_{1 \leqslant i \leqslant n} X_i \leqslant B + c\Big] = \mathrm{o}\big(\mathbb{P}[A_e \geqslant B]\big),$$

where the last asymptotic relation follows from lemma A.4. This completes the proof of (i).

To prove (ii) let us construct a sequence $\{Y_k, \ k \geqslant 1\}$ of integer valued i.i.d. random variables in the following way. First, choose an integer n_0 such that $\sum_{i=n_0}^{\infty} K\mathbb{P}[A \geqslant i - c] < 1$. Then, assign to each Y_k the following distribution:

$$\mathbb{P}[Y_k \geqslant i] = \begin{cases} 1 & \text{for } i \leqslant n_0, \\ \sum_{n=i}^{\infty} K\mathbb{P}[A \geqslant n - c] & \text{otherwise.} \end{cases}$$

Clearly, Y_k is stochastically larger than X_k^B for all B, i.e.,

$$\mathbb{P}\big[X_k^B \geqslant i\big] \leqslant \mathbb{P}[Y_k \geqslant i],$$

and $Y_k \in \mathcal{S}$. Thus,

$$\mathbb{P}\big[S_n^B \geqslant B\big] \leqslant \mathbb{P}\left[\sum_{i=1}^{n} Y_k \geqslant B\right] \leqslant K_\varepsilon (1+\varepsilon)^n \mathbb{P}[A_e \geqslant B],$$

where the last inequality follows from lemma 2.10 of the Appendix of this special issue of Queueing Systems. This concludes the proof of the lemma. □

Proof of lemma 2. Finally, we are ready to provide the proof of lemma 2. Recall that the proof will follow if we show that (A.28) holds. First, observe that $g_+^B(1) = \mathbb{P}[Q^{\infty,B} > 0]$ monotonically increases to $g_+^{\infty}(1) = \mathbb{P}[Q^{\infty,\infty} > 0]$ as $B \to \infty$, where $Q^{\infty,\infty}$ is the workload of an infinite buffer queue with (non-truncated) arrival sequence A_n. Hence, for any δ_2 such that $0 < \delta_2 < g_+^{\infty}(1) < g_+^{\infty}(1) + \delta_2 < 1$, we can choose B_0 such that for all $B \geqslant B_0$

$$\delta_2 < g_+^B(1) < g_+^{\infty}(1) + \delta_2 < 1. \tag{A.38}$$

Next, let us choose a distribution for random variables in lemma A.5 to be $\mathbb{P}[X_i^B = k] = g_{+k}^B / g_+^B(1)$, $1 \leqslant k \leqslant B + c$. By lemma A.3 and (A.38) it follows that for all $B \geqslant B_0$

$$\mathbb{P}\big[X_i^B = k\big] \leqslant \frac{K_3}{\delta_2} \mathbb{P}[A \geqslant k - c],$$

where K_3 is the same as in lemma A.3. Thus, the condition of lemma A.5 is satisfied.

Then, from equation (A.30), similarly as in (A.2), $\mathbb{P}[Q^{\infty,B} > B]$ can be represented as

$$\mathbb{P}\big[Q^{\infty,B} > B\big] = \big(1 - g_+^B(1)\big) \sum_{n=1}^{\infty} \big(g_+^B(1)\big)^n \mathbb{P}\big[S_n^B > B\big]. \tag{A.39}$$

Hence, by (A.38), for all sufficiently large B

$$\mathbb{P}\big[Q^{\infty,B} > B\big] \leqslant \sum_{n=1}^{\infty} \big(g_+^{\infty}(1) + \delta_2\big)^n \mathbb{P}\big[S_n^B > B\big]. \tag{A.40}$$

Thus, by applying lemma A.5 (with $\varepsilon > 0$ such that $(1+\varepsilon)(g_+^{\infty}(1) + \delta_2) < 1$) and the Dominated Convergence theorem to the sum in (A.40) we derive

$$\mathbb{P}\big[Q^{\infty,B} > B\big] = \mathrm{o}\big(\mathbb{P}[A_e > B\big) \quad \text{as } B \to \infty,$$

which concludes the proof of the lemma 2.

References

[1] J. Abate, G.L. Choudhury and W. Whitt, Waiting-time tail probabilities in queues with long-tail service-time distributions. Queueing Systems 16(3/4) (1994) 311–338.

[2] V. Anantharam, On the sojourn time of sessions at an ATM buffer with long-range dependent input traffic, in: *Proceedings of the 34th IEEE Conference on Decision and Control* (December 1995).

[3] S. Asmussen, *Applied Probability and Queues* (Wiley, 1987).

[4] S. Asmussen, L.F. Henriksen and C. Klüppelberg, Large claims approximations for risk processes in a Markovian environment, Stochastic Process. Appl. 54 (1994) 29–43.

[5] S. Asmussen, H. Schmidli and V. Schmidt, Tail probabilities for non-standard risk and queueing processes with subexponential jumps, Preprint (1997).

[6] N.H. Bingham, C.M. Goldie and J.L. Teugels, *Regular Variation* (Cambridge University Press, Cambridge, 1987).

[7] A.A. Borovkov, *Asymptotic Methods in Queueing Theory* (Wiley, 1984).

[8] O.J. Boxma, Regular variation in a multi-source fluid queue, in: *Proc 15th ITC*, Washington, DC (June 1997) pp. 391–402.

[9] O.J. Boxma and V. Dumas, Fluid queues with long-tailed activity period distributions, Technical Report PNA-R9705, CWI, Amsterdam (April 1997).

[10] G.L. Choudhury and W. Whitt, Long-tail buffer-content distributions in broadband networks, Performance Evaluation 30 (1997) 177–190.

[11] J. Chover, P. Ney and S. Wainger, Functions of probability measures, J. Anal. Math. 26 (1973) 255–302.

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

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

[14] K. Debicki, Z. Michna and T. Rolski, On the supremum from Gaussian processes over infinite horizon, Preprint (May 1997).

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

[16] P. Embrechts and J. Hawkes, A limit theorem for the tails of discrete infinitely divisable laws with applications to fluctuation theory, J. Austral. Math. Soc. Ser. A 32 (1982) 412–422.

[17] W. Feller, *An Introduction to Probability Theory and its Application*, Vol. II (Wiley, New York, 1971).

[18] C.M. Goldie and C. Klüppelberg, Subexponential distributions, in: *A Practical Guide to Heavy Tails: Statistical Techniques for Analysing Heavy Tailed Distributions*, eds. M.S. Taqqu, R. Adler and R. Feldman (Birkhäuser, Boston, 1998) pp. 435–459.

[19] J.J. Gordon, K. Murti and A. Rayes, Overview of Internet traffic issues on the PSTN, in: *Proc. 15th ITC*, Washington, DC (June 1997) pp. 643–652.

[20] D. Heath, S. Resnick and G. Samorodnitsky, Patterns of buffer overflow in a class of queues with long memory in the input stream, Ann. Appl. Probab. 7 (1997) 1021–1057.

[21] D. Heath, S. Resnick and G. Samorodnitsky, Heavy tails and long range dependence in on/off processes and associated fluid models, Math. Oper. Res. 23 (1998) 145–165.

[22] D.P. Heyman and T.V. Lakshman, Source models for VBR broadcast-video traffic, IEEE/ACM Trans. Networking 4 (February 1996) 40–48.

[23] P.R. Jelenković, Long-tailed loss rates in single server queue, Technical Report BL0112120-970808-TM09, Bell Labs, Lucent Technologies (1997).

[24] P.R. Jelenković and A.A. Lazar, Subexponential asymptotics of a Markov-modulated random walk with queueing applications, J. Appl. Probab. 35(2) (June 1998) 325–347.

[25] P.R. Jelenković and A.A. Lazar, Asymptotic results for multiplexing subexponential on–off processes, Adv. Appl. Probab. 31(2) (1999, to appear).

[26] P.R. Jelenković, A.A. Lazar and N. Semret, The effect of multiple time scales and subexponentiality of MPEG video streams on queueing behavior, IEEE J. Selected Areas Commun. 15(6) (August 1997) 1052–1071.

[27] C. Klüppelberg, Subexponential distributions and integrated tails, J. Appl. Probab. 25 (1988) 132–141.

[28] C. Klüppelberg, Subexponential distributions and characterizations of related classes, Probab. Theory Related Fields 82 (1989) 259–269.

[29] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, On the self-similar nature of Ethernet traffic, in: *SIGCOMM '93* (1993) pp. 183–193.

[30] N. Likhanov, B. Tsybakov and N.D. Georganas, Analysis of an ATM buffer with self-similar ("fractal") input traffic, in: *INFOCOM '95*, Boston, MA (April 1995) pp. 985–991.

[31] R.M. Loynes, The stability of a queue with non-independent inter-arrival and service times, Proc. Cambridge Philos Soc. 58 (1968) 497–520.

[32] P. Nain, Z. Liu, D. Towsley and Z.-L. Zhang, Asymptotic behavior of a multiplexer fed by a long-range dependent process, Preprint (June 1997).

[33] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach* (Johns Hopkins University Press, Baltimore, MD, 1981).

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

[35] A.G. Pakes, On the tails of waiting-time distribution, J. Appl. Probab. 12 (1975) 555–564.

[36] M. Parulekar and A.M. Makowski, Tail probabilitites for a multiplexer with self-similar traffic, in: *INFOCOM '96*, San Francisco, CA (March 1996).

[37] M. Parulekar and A.M. Makowski, Tail probabilitites for $M/G/\infty$ input processes (I): preliminary asymptotices, Queueing Systems 27 (1997) 271–296.

[38] S. Resnick and G. Samorodnitsky, Performance decay in a single server queueing model with long range dependence, Preprint (1996).

[39] S. Resnick and G. Samorodnitsky, Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues, Preprint (1997).

[40] T. Rolski, S. Schlegel and V. Schmidt, Asymptotics of palm-stationary buffer content distribution in fluid flow queues, Preprint (1997).

[41] B.K. Ryu and S.B. Lowen, Point process approaches to the modeling and analysis of self-similar traffic — part I: Model construction, in: *INFOCOM '96*, San Francisco, CA (March 1996).
[42] K.P. Tsoukatos and A.M. Makowski, Heavy traffic analysis for a multiplexer driven by M/GI/∞ input processes, in: *Proc. 15th ITC*, Washington, DC (June 1997) pp. 497–506.
[43] N. Veraverbeke, Asymptotic behavior of Wiener–Hopf factors of a random walk, Stochastic Process. Appl. 5 (1977) 27–37.
[44] E. Willekens and J.L. Teugels, Asymptotic expansion for waiting time probabilities in an M/G/1 queue with long-tailed service time, Queueing Systems 10 (1992) 295–312.
[45] A.P. Zwart, A fluid queue with a finite buffer and subexponential input, Preprint (1999).