ASYMPTOTIC BEHAVIOR OF A MULTIPLEXER FED BY A LONG-RANGE DEPENDENT PROCESS*

Zhen LIU[1] Philippe NAIN[1] Don TOWSLEY[2] and Zhi-Li ZHANG[3]

[1]INRIA
B.P. 93, 06902, Sophia Antipolis Cedex, France
{liu, nain}@sophia.inria.fr

[2]Department of Computer Science
University of Massachusetts, Amherst, MA 01003, USA
towsley@cs.umass.edu

[3]Department of Computer Science and Engineering
University of Minnesota, 200 Union St. S.E., Minneapolis, MN 55455, USA
zhzhang@cs.umn.edu

Journal of Applied Probability, Vol. 26, pp. 105-118, 1999

Abstract

In this paper we study the asymptotic behavior of the tail of the stationary backlog distribution in a single server queue with constant service capacity c, fed by the so-called "$M/G/\infty$ input process" or "Cox input process". Asymptotic lower bounds are obtained for any distribution G and asymptotic upper bounds are derived when G is a subexponential distribution. We find the bounds to be tight in some instances, e.g., G corresponding to either the Pareto or lognormal distribution and $c - \rho < 1$, where ρ is the arrival rate to the buffer.

Keywords: Asymptotic self-similar process; Long-range dependence; Subexponential distributions; Pareto distribution; Large deviations; Queues.

*This work was performed in part while D. Towsley and Z.-L. Zhang were visiting INRIA. It was also funded in part by the National Science Foundation under grants NCR-9508274, NCR-9623807, and CDA 9502639. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

1 Introduction

The recent discovery [16, 21, 30] that traffic in networks possess long-range time dependencies that cannot be easily captured by Poisson-based models has motivated queueing theorists to propose and analyze new queueing models that capture these dependencies. One such model that has received attention is a buffer with server having rate c fed by an $M/G/\infty$ input process where G is heavy-tailed (e.g., [1, 12, 19, 27]). This is of interest because of its versatility, i.e., the dependencies over different time-scales can be controlled by varying the tail behavior of G.

In this paper we consider the model introduced by Parulekar and Makowski [27]. A discrete-time single-server queue (called the multiplexer) with infinite waiting room and with service capacity c is fed by an integer-valued process $\{b_t,\, t \in \mathbb{N}\}$. The r.v. b_t is defined as the number of busy servers at time $t \in \mathbb{N}$ in an M/G/∞ queue with arrival intensity $\lambda > 0$ and i.i.d. service times $\{\sigma_n\}_n$ with common cumulative distribution function (c.d.f.) $G(x) = P(\sigma_n \leq x)$ and finite mean $\overline{\sigma}$. An appealing feature of the (stationary version of the) input process $\{b_t,\, t \in \mathbb{N}\}$ is that it is a long-range dependent process [2] for some well-chosen *subexponential* c.d.f.'s G (see Section 2).

Let Q_t be the queue-length at the multiplexer at time t. Then, Q_t satisfies the Lindley's equation $Q_{t+1} = \max(0, Q_t + b_t - c)$ for all $t \in \mathbb{N}$, with $Q_0 = 0$. Let Q be the stationary queue-length under the stability condition $c > \rho := \lambda\overline{\sigma}$ (see Section 2). The aim of this paper is to study the behavior of $\log P(Q > x)$ and of $P(Q > x)$ for large x. More precisely, we show that there exist positive and finite numbers θ_1 and θ_2, depending on G, such that

$$-\theta_1 \leq \liminf_{x\to\infty} \frac{\log P(Q > x)}{-\log \overline{G_1}(x)} \leq \limsup_{x\to\infty} \frac{\log P(Q > x)}{-\log \overline{G_1}(x)} \leq -\theta_2. \tag{1}$$

The lower bound in (1) holds for any c.d.f. G whereas the upper bound holds for any *subexponential* c.d.f. G (to be defined in Section 2). Here G_1 is defined as

$$G_1(x) := \frac{1}{\overline{\sigma}} \int_0^x \overline{G}(u)\, du, \quad x \geq 0 \tag{2}$$

and $\overline{F}(x) = 1 - F(x)$ for any probability distribution F. We also show that the bounds in (1) are tight (i.e. $\theta_1 = \theta_2$) when G is Pareto or lognormal (see Corollary 4.1), provided that $c - \rho < 1$. In the following the bounds in (1) will be referred to as *large deviations* bounds. Asymptotic upper and lower bounds for $P(Q > x)$ are also obtained.

Large deviations bounds were obtained in [29] in the case when G is short-tailed. Duffield observed in [12] that the approach in [27], based on the Gärtner-Ellis theorem, cannot be used to derive large deviations *lower* bounds for heavy-tailed G. By refining Theorem 2.2 in [13] and by using results in [28] Duffield was able to obtain the following large deviations *upper* bound (see [12])

$$\limsup_{x\to\infty} \frac{\log P(Q > x)}{\log x} \leq 1 - (\alpha - 1)(c - \rho) \tag{3}$$

in the case of the Pareto distribution $\overline{G}(x) \sim c_1 x^{-\alpha}$. An asymptotic lower bound for $P(Q > x)$ was obtained by Jelenkovic and Lazar [19] in the case when $c - \rho < 1$ and under a technical condition on G_1 (see comment after the proof of Proposition 3.2).

In this paper we propose an alternative to the approach based on the Gärtner-Ellis theorem that will yield asymptotic lower and upper bounds. We will observe that the large deviations bounds are tight for a number of subexponential distributions when $c - \rho < 1$ and that, in the case of G Pareto, the large deviations upper bound that can be derived from (1) (see Proposition 4.1) is tighter than that of Duffield when $c - \rho \leq \alpha/(\alpha - 1)$; otherwise Duffield's is tighter.

Other models have been proposed for modeling the effects of long-range dependence in arrival processes on buffer occupancy statistics. These include fractional brownian motion [13, 25], fractional gaussian noise [27], and a finite population of on-off sources where the on state holding times are characterized by heavy-tailed distributions [5, 7, 9, 18, 19, 22, 31] (see [6] for a survey on fluid queues with long-tailed activity periods).

The rest of the paper is structured as follows. Section 2 contains a characterization of the stationary behavior of the $M/G/\infty$ input process and the definition and characterization of the family of subexponential distributions. Asymptotic lower and upper bounds are established in Sections 3 and 4 respectively. Concluding remarks on the superposition of independent M/G/∞ input processes are given in Section 5.

2 Preliminaries

The lemma below gives a useful characterization of the stationary behavior of the input process $\{b_t,\, t \in \mathbb{N}\}$. We will assume that customers entering the M/G/∞ queue begin their service upon arrival (see Remark 2.1).

Lemma 2.1 *The distribution of the sequence $\{b_{t+k},\, t \in \mathbb{N}\}$ converges monotonically for $k \to \infty$ to that of a proper stationary and ergodic sequence $\{b^t,\, t \in \mathbb{N}\}$ such that*

$$b^t \stackrel{\text{st}}{=} \sum_{j=0}^{b^0} I\left(\hat{\sigma}_j > t\right) + \sum_{s=0}^{t-1} \sum_{s \leq T_j < s+1} I(\sigma_j > t - T_j), \quad t \in \mathbb{N} \tag{4}$$

where

(i) $0 < T_1 \leq T_2 \leq \cdots$ *are the successive jump times of a Poisson process with intensity* λ, *independent of the service times* $\{\sigma_n,\, n = 1, 2, \ldots\}$;

(ii) b^0 *is a Poisson r.v. with parameter* $\rho := \lambda\, \overline{\sigma}$;

(iii) conditioned on the event $\{b^0 = k\}$, $k \geq 1$, the r.v.'s $\{\hat{\sigma}_1, \ldots, \hat{\sigma}_k\}$ are i.i.d. with common c.d.f. G_1 as defined in (2), namely,

$$P\left(\hat{\sigma}_1 \leq x_1, \ldots, \hat{\sigma}_k \leq x_k \,|\, b^0 = k\right) = \prod_{j=1}^{k} G_1(x_j).$$

Further, the r.v.'s $\{T_j, \sigma_j,\, j = 1, 2, \ldots\}$ are independent of the r.v.'s $\{b^0, \hat{\sigma}_j,\, j = 1, 2, \ldots\}$.

The proof of this lemma follows from [4, Chapter 6] and [33, pp. 160-162] (see also [27]). The interpretation of (4) is the following: given that the $M/G/\infty$ queue is in steady-state at time $t = 0$, the first sum in the r.h.s. gives the number of busy servers at time $t = 1, 2\ldots$ among all servers busy at time $0-$; the second sum gives the number of servers that became busy at time s, $0 \leq s \leq t-1$, and that are still busy at time t.

Assume that $\rho < c$. Since the process $\{b_{t+k},\, t \in N\}$ converges to the stationary and ergodic process $\{b^t,\, t \in \mathbb{N}\}$ (see Lemma 2.1) then it is well-known (see e.g. [4, Theorem 6, p. 12]) that there exists a proper r.v. Q such that

$$P(Q > x) = \lim_{t \to \infty} P(Q_t > x) = P\left(\sup_{t \in \mathbb{N}} \left(\sum_{s=0}^{t-1} b^{-s} - ct\right) > x\right), \quad x \in \mathbb{N} \tag{5}$$

where $\{b^t,\, -\infty < t < \infty\}$ is a stationary and ergodic process obtained by supplementing $\{b^t,\, t \in \mathbb{N}\}$. We will however prefer the following representation for the stationary queue length distribution:

$$P(Q > x) = P\left(\sup_{t \in \mathbb{N}} \left(\sum_{s=0}^{t-1} b^{s} - ct\right) > x\right), \quad x \in \mathbb{N}, \tag{6}$$

which follows from (5) together with the property that the number of busy servers in a stationary M/G/∞ queue is a reversible stochastic process [20, Theorem 3.11].

The rest of this paper is devoted to the computation of asymptotic lower and upper bounds for $P(Q > x)$. Particular attention will be devoted to the case when the c.d.f. G of the service times is *subexponential*. Recall that a probability distribution F on $[0, \infty)$ is subexponential, denoted as $F \in \mathcal{S}$ (or $\overline{F} \in \mathcal{S}$ with a slight abuse of notation) if $\overline{F^{\star 2}}(x) \sim 2\,\overline{F}(x)$ where $F^{\star 2}$ denotes the 2nd convolution of F with itself, namely, $F^{\star 2}(x) = \int_0^\infty F(x-u)\, F(du)$. As usual, the notation $f(x) \sim g(x)$ will stand for $\lim_{x\to\infty} f(x)/g(x) = 1$ and $f(x) = o(g(x))$ will stand for $\lim_{x\to\infty} f(x)/g(x) = 0$. The class of subexponential distributions was introduced by Chistakov [8] and contains Pareto, Weibull and lognormal distributions (see Section 3), among others. A probability distribution F on $[0, \infty)$ belongs to the class $\mathcal{D}$ of dominated-variation distributions if $\limsup_{x\to\infty} \overline{F}(x)/\overline{F}(2x) < \infty$ and to the class $\mathcal{L}$ of long-tailed distributions if $\lim_{x\to\infty} \overline{F}(x-y)/\overline{F}(x) = 1$ for all $y \in (-\infty, \infty)$.

For any c.d.f. F on $[0,\infty)$ with finite expectation μ, (i.e. $\mu := \int_0^\infty u\, F(du) < \infty$), define the integrated tail distribution F_1 by

$$F_1(x) := \frac{1}{\mu}\int_0^x \overline{F}(u)\, du, \quad x \geq 0.$$

Note that G_1 in (2) is the integrated tail distribution of σ_n.

The next lemma reports basic properties of subexponential probability distributions.

Lemma 2.2 *The following statements hold:*

(a) $\mathcal{D} \cap \mathcal{L} \subset \mathcal{S} \subset \mathcal{L}$ **[15, 17]**;

(b) If F has finite expectation and if $F \in \mathcal{D}$ then $F_1 \in \mathcal{D} \cap \mathcal{L}$ **[15]**;

(c) If $F \in \mathcal{S}$ and G is a probability distribution on $[0,\infty)$ such that $\overline{F}(x) \sim c_1\, \overline{G}(x)$ for some positive constant c_1, then $G \in \mathcal{S}$ **[26, Lemma 2]**.

In particular, we see from properties (a) and (b) that if $F \in \mathcal{D} \cap \mathcal{L}$ and if F has finite expectation then $F, F_1 \in \mathcal{S}$.

We conclude this section by pointing out an interesting feature (already observed in [27, p. 1455]) of the process $\{b^t,\, t \in \mathbb{N}\}$ defined in (4). First, it has been shown in [11, formula (5.39)] that $\mathrm{cov}(b^t, b^{t+h}) = \rho\, \overline{G_1}(h)$ for all $t, h \in \mathbb{N}$. Therefore, the stationary process $\{b^t,\, t \in \mathbb{N}\}$ will be long-range dependent [2] if $\sum_{h=0}^\infty \overline{G_1}(h) = \infty$, which will occur, for instance, when G is Pareto (i.e. $\overline{G}(x) \sim x^{-\alpha}$) with parameter $1 < \alpha < 2$.

Remark 2.1 *By taking integer-valued service times our model reduces to that in [27]. This follows from the fact that in the case of integer-valued service times the number of busy servers at time $t+1$ is the same whether customers entering the $M/G/\infty$ queue in $(t, t+1)$ begin their service upon arrival (as in our model) or begin their service at time $t+1$ (as in [27]).*

3 Lower Bounds

The following representation of $A(0,t) := \sum_{s=0}^{t-1} b^s$ will prove useful:

$$\begin{aligned} A(0,t) &= \sum_{s=0}^{t-1} b^s \\ &= \sum_{s=0}^{t-1} \sum_{j=1}^{b^0} I\,(\hat{\sigma}_j > s) + \sum_{s=0}^{t-1} \sum_{k=0}^{s-1} \sum_{k \leq T_j < k+1} I(\sigma_j > s - T_j) \end{aligned}$$

$$= \sum_{j=1}^{b^0} \sum_{s=0}^{t-1} I(\hat{\sigma}_j > s) + \sum_{k=0}^{t-2} \sum_{k \le T_j < k+1} \sum_{s=k+1}^{t-1} I(\sigma_j > s - T_j)$$

$$= \sum_{j=1}^{b^0} \min(\lceil \hat{\sigma}_j \rceil, t) + \sum_{k=0}^{t-2} \sum_{k \le T_j < k+1} \sum_{s=k+1}^{t-1} I(\sigma_j > s - T_j) \tag{7}$$

where $\lceil x \rceil$ denotes the smallest integer larger than or equal to x.

The first sum in the r.h.s. of (7) gives the total number of customers arriving to the multiplexer in $[0, t)$ generated by all servers in the infinite-server queue busy at time 0; the second sum gives the total number of customers arriving to the multiplexer in $(0, t)$ generated by all servers in the infinite-server queue that become active at time $1, 2, \ldots, t-1$. Set

$$a_0(t) := \sum_{j=1}^{b^0} \min(\lceil \hat{\sigma}_j \rceil, t) \tag{8}$$

$$a_s(t) := \sum_{s-1 \le T_j < s} \sum_{i=s}^{t-1} I(\sigma_j > i - T_j), \quad s = 1, 2, \ldots, t-1 \tag{9}$$

so that

$$A(0, t) = \sum_{s=0}^{t-1} a_s(t). \tag{10}$$

Denote by $\lfloor x \rfloor$ the largest integer smaller than or equal to x. The following asymptotic lower bound for $\log P(Q > x)$ holds:

Proposition 3.1 (Large deviations lower bound)

For any c.d.f. G,

$$\liminf_{x \to \infty} \frac{\log P(Q > x)}{-\log \overline{G_1}(x)} \ge -\inf_{\beta > 0} \left\{ (\lfloor c - \rho + \beta \rfloor + 1) \limsup_{x \to \infty} \frac{\log \overline{G_1}(x)}{\log \overline{G_1}(\beta x)} \right\}. \tag{11}$$

Proof. Fix $\beta > 0$, $\epsilon > 0$, and define $\gamma := c - \rho + \beta + \epsilon$. Note that $\gamma > 0$ under the stability condition $c > \rho$.

We have

$$\liminf_{x \to \infty} \frac{\log P(Q > x)}{-\log \overline{G_1}(x)} = \liminf_{t \to \infty} \frac{\log P(Q > \beta t)}{-\log \overline{G_1}(\beta t)}$$

$$\ge \liminf_{t \to \infty} \frac{\log P(A(0,t) - ct > \beta t)}{-\log \overline{G_1}(\beta t)} \tag{12}$$

$$\ge \liminf_{t \to \infty} \frac{-1}{\log \overline{G_1}(\beta t)} \log P\left(a_0(t) \ge \gamma t, \sum_{s=1}^{t-1} a_s(t) > (\rho - \epsilon) t \right)$$

$$= \liminf_{t\to\infty} \frac{-1}{\log \overline{G_1}(\beta t)} \left[\log\, P(a_0(t) \geq \gamma t) + \log P\left(\sum_{s=1}^{t-1} a_s(t) > (\rho-\epsilon)t\right)\right] \tag{13}$$

$$\geq \liminf_{t\to\infty} \frac{\log\, P(a_0(t)\geq \gamma t)}{-\log \overline{G_1}(\beta t)} + \liminf_{t\to\infty} \frac{-1}{\log \overline{G_1}(\beta t)}\, \log\, P\left(\sum_{s=1}^{t-1} a_s(t) > (\rho-\epsilon)t\right). \tag{14}$$

Inequality (12) follows from $P(Q>x) \geq P(A(0,t)-ct>x)$ (see (6)); (13) is a consequence of the independence of the r.v.'s $a_0(t)$ and $\sum_{s=1}^{t} a_s(t)$ (see Lemma 2.1); (14) comes from the inequality $\liminf_n (a_n+b_n) \geq \liminf_n a_n + \liminf_n b_n$.

Let us now focus on the first limit in the r.h.s. of (14). We have for $t>0$

$$\begin{aligned} P(a_0(t)\geq\gamma t) &= P\left(\sum_{j=1}^{b^0} \min\left(\lceil\hat\sigma_j\rceil, t\right) \geq \gamma t\right) \\ &\geq \sum_{k=\lceil\gamma\rceil}^{\infty} P\left(\sum_{j=1}^{k} \min(\hat\sigma_j, t) \geq \gamma t \,|\, b^0 = k\right) P(b^0=k) && (15) \\ &\geq \sum_{k=\lceil\gamma\rceil}^{\infty} P\left(\hat\sigma_1 > t, \cdots, \hat\sigma_{\lceil\gamma\rceil} > t \,|\, b^0=k\right) P(b^0=k) \\ &= \overline{G_1}(t)^{\lceil\gamma\rceil}\, P(b^0 \geq \lceil\gamma\rceil) && (16) \end{aligned}$$

where (16) follows from Lemma 2.1(iii).

Since $P(b^0\geq\lceil\gamma\rceil)>0$ (see Lemma 2.1(ii)) we deduce from (16) that

$$\liminf_{t\to\infty} \frac{\log P(a_0(t)\geq\gamma t)}{-\log\overline{G_1}(\beta t)} \geq -\lceil\gamma\rceil\, \limsup_{t\to\infty} \frac{\log\overline{G_1}(t)}{\log\overline{G_1}(\beta t)}. \tag{17}$$

Let us show that the second limit in the r.h.s. of (14) is 0. We see from the definition of $A(0,t)$ and from (8)-(10) that

$$\sum_{s=1}^{t-1} a_s(t) \geq \sum_{s=0}^{t-1} b^s - \sum_{j=1}^{b^0} \lceil\hat\sigma_j\rceil. \tag{18}$$

On the other hand, the stationarity and ergodicity of the sequence $\{b^t,\, t\in\mathbb{N}\}$ together with $\rho = E[b^0] < \infty$ (see Lemma 2.1) yields

$$\lim_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} b^s = \rho \quad \text{a.s.} \tag{19}$$

from ergodic theory (see e.g. [32, Chapter V]). We therefore deduce from (18)-(19) that

$$\liminf_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t-1} a_s(t) \geq \rho \quad \text{a.s.} \tag{20}$$

since $\sum_{j=1}^{b^0} \hat{\sigma}_j < \infty$ a.s. by Lemma 2.1.

Combining [24, Proposition I-4-3] together with (20) yields

$$1 \geq \liminf_t P\left(\sum_{s=1}^{t-1} a_s(t) > (\rho - \epsilon)t\right) \geq P\left(\liminf_t \left\{\sum_{s=1}^{t-1} a_s(t) > (\rho - \epsilon)t\right\}\right) = 1 \qquad (21)$$

which entails that

$$\liminf_{t\to\infty} \frac{-1}{\log \overline{G_1}(\beta t)} \log P\left(\sum_{s=1}^{t-1} a_s(t) > (\rho - \epsilon)t\right) = 0. \qquad (22)$$

In summary, we have shown that (cf. (14), (17), (22))

$$\begin{aligned}\liminf_{x\to\infty} \frac{\log P(Q > x)}{-\log \overline{G_1}(x)} &\geq -\inf_{\beta>0,\epsilon>0} \left\{\lceil c - \rho + \beta + \epsilon\rceil \limsup_{t\to\infty} \frac{\log \overline{G_1}(t)}{\log \overline{G_1}(\beta t)}\right\} \\ &\geq -\inf_{\beta>0} \left\{(\lfloor c - \rho + \beta\rfloor + 1) \limsup_{t\to\infty} \frac{\log \overline{G_1}(t)}{\log \overline{G_1}(\beta t)}\right\}\end{aligned}$$

which completes the proof. ∎

It is worth noting that the lower bound in (11) is never trivial as it is always larger than or equal to $-(\lfloor c - \rho\rfloor + 2)$ that is obtained for $\beta = 1$.

The next result proposes asymptotic lower bounds for $P(Q > x)$.

Proposition 3.2 (Asymptotic lower bound)

For any c.d.f. G,

$$\liminf_{x\to\infty} \frac{P(Q > x)}{\overline{G_1}(x)^{\lfloor c-\rho\rfloor+1}} \geq \sup_{0<\beta<1+\lfloor c-\rho\rfloor-(c-\rho)} \liminf_{x\to\infty} \left(\frac{\overline{G_1}(x)}{\overline{G_1}(\beta x)}\right)^{\lfloor c-\rho\rfloor+1} \left(1 - \sum_{k=0}^{\lfloor c-\rho\rfloor} \frac{\rho^k}{k!} e^{-\rho}\right). \qquad (23)$$

Proof. The proof of (23) follows the same line of arguments as that of Proposition 3.1. Define $\gamma := c - \rho + \beta + \epsilon$. Let $0 < \beta < 1 + \lfloor c - \rho\rfloor - (c - \rho)$ and pick $\epsilon > 0$ small enough so that $\lceil\gamma\rceil = \lfloor c - \rho\rfloor + 1$.

In direct analogy with the derivation of (14) and by using (16) and (21) we get

$$\liminf_{x\to\infty} \frac{P(Q > x)}{\overline{G_1}(x)^{\lfloor c-\rho\rfloor+1}} \geq \liminf_{t\to\infty} \frac{P(a_0(t) > \gamma t)}{\overline{G_1}(\beta t)^{\lfloor c-\rho\rfloor+1}} \qquad (24)$$

$$\geq \liminf_{t\to\infty} \left(\frac{\overline{G_1}(t)}{\overline{G_1}(\beta t)}\right)^{\lfloor c-\rho\rfloor+1} P(b^0 \geq \lfloor c - \rho\rfloor + 1) \qquad (25)$$

for all $0 < \beta < 1 + \lfloor c - \rho \rfloor - (c - \rho)$, from which (23) follows. ∎

It is worth noting that the supremum in the r.h.s. of (23) is strictly positive if and only if $G_1 \in \mathcal{D}$. Indeed, it follows from [3, Corollary 2.0.6, p. 65] that if $\liminf_{x\to\infty} \overline{G_1}(x)/\overline{G_1}(\delta x)$ is strictly positive for some $\delta \in (0,1)$ then this limit is strictly positive for all $\delta \in (0,1)$, and in particular for $\delta = 1/2$. A sufficient condition for $G_1 \in \mathcal{D}$ is that $G \in \mathcal{D}$ (e.g. G Pareto) and G has finite expectation (see Lemma 2.2(b)).

A refined lower bound has been obtained in [23] under the additional assumption that $G_1 \in \mathcal{S}$. When $c - \rho < 1$, Jelenkovic and Lazar [19, Theorem 11] have derived a tighter lower bound with the same decay function $\overline{G_1}(x)$ but with a larger coefficient. The bound in [19] holds provided that $L := \lim_{\delta\downarrow 1} \liminf_{x\uparrow\infty} \overline{G_1}(\delta x)/\overline{G_1}(x) > 0$ (Jelenkovic and Lazar [19] actually assume that $L = 1$ but this assumption can be weakened to $L > 0$; if so, then the coefficient of their lower bound in Theorem 11 has to be multiplied by L). Since $\overline{G_1}$ is non-increasing, it is easy to see from [3, Corollary 2.0.6] that $L > 0$ is equivalent to $G_1 \in \mathcal{D}$. Hence, both bounds in Proposition 3.2 and in [19] are non-trivial if and only if $G_1 \in \mathcal{D}$.

Corollary 3.1 *When $G_1 \in \mathcal{D}$ then*

$$\liminf_{x\to\infty} \frac{\log P(Q > x)}{-\log \overline{G_1}(x)} \geq -\lfloor c - \rho \rfloor - 1. \tag{26}$$

When Corollary 3.1 applies, the lower bound in the r.h.s. of (26) is easier to compute than the lower bound in Proposition 3.1 but may not be as tight (for G Pareto both bounds in (11) and in (26) are the same as reported below).

We conclude this section by addressing the cases when G is (i) geometric, (ii) Pareto, (iii) Weibull, and (iv) lognormal.

(i) G is geometric. We have $P(\sigma_n = r) = (1-q)\,q^{r-1}$ for $r = 1, 2, \ldots$ with $q \in (0,1)$. Hence, $\overline{G_1}(r) = q^r$ for $r = 0, 1, \ldots$. Proposition 3.2 yields a trivial lower bound ($= 0$). From Proposition 3.1 we find

$$\liminf_{x\to\infty} \frac{1}{x} \log P(Q > x) \geq \log q \inf_{\beta>0} \frac{\lfloor c - \rho + \beta \rfloor + 1}{\beta} = \log q. \tag{27}$$

The r.h.s. of (27) follows from the inequalities

$$\frac{c - \rho + \beta + 1}{\beta} \geq \frac{\lfloor c - \rho + \beta \rfloor + 1}{\beta} \geq 1$$

together with $\lim_{\beta\to\infty} (c - \rho + \beta + 1)/\beta = 1$.

(ii) G is Pareto. We have $\overline{G}(x) \sim c_1\, x^{-\alpha}$ for some $\alpha > 1$, $c_1 > 0$. Hence,

$$\overline{G_1}(x) \sim c_2\, x^{-\alpha+1} \tag{28}$$

with $c_2 = c_1/(\overline{\sigma}(\alpha-1))$. From Proposition 3.2 we get

$$\liminf_{x\to\infty} \frac{P(Q>x)}{x^{(-\alpha+1)\,\zeta}} \geq c_2^{\zeta}\,(\zeta-(c-\rho))^{(\alpha-1)\,\zeta}\left(1-\sum_{k=0}^{\zeta-1}\frac{\rho^k}{k!}\,e^{-\rho}\right). \tag{29}$$

where we set $\zeta := \lfloor c-\rho\rfloor + 1$. In particular, (29) (or Proposition 3.1/ Corollary 3.1) yields

$$\liminf_{x\to\infty} \frac{1}{\log x}\,\log\, P(Q>x) \geq (-\alpha+1)\,\zeta. \tag{30}$$

(iii) G is Weibull. We have $\overline{G}(x) = e^{-c_1 x^{\nu}}$ for some $0<\nu<1$ and $c_1>0$. Simple algebra yield

$$\overline{G_1}(x) \sim c_2\, e^{-c_1 x^{\nu}}\, x^{1-\nu} \tag{31}$$

with $c_2 = 1/(c_1\nu\overline{\sigma})$ and $\overline{\sigma} = \Gamma(1/\nu)/(\nu c_1^{1/\nu})$ where $\Gamma(s) := \int_0^\infty x^{s-1}\exp(-x)dx$ for $s>0$. Proposition 3.2 yields a trivial lower bound (i.e. 0). By Proposition 3.1 we get (Corollary 3.1 does not apply since $G_1 \notin \mathcal{D}$)

$$\begin{aligned}\liminf_{x\to\infty} \frac{1}{x^{\nu}}\,\log\, P(Q>x) &\geq -\inf_{\beta>0}\frac{\lfloor c-\rho+\beta\rfloor+1}{\beta^{\nu}}\\ &= \begin{cases} -\min\left\{\dfrac{\lfloor c-\rho\rfloor+\lfloor a\rfloor}{(\lfloor a\rfloor-q)^{\nu}};\dfrac{\lfloor c-\rho\rfloor+\lceil a\rceil}{(\lceil a\rceil-q)^{\nu}}\right\}, & \text{if } a\geq 1\\[2ex] -\dfrac{\lfloor c-\rho\rfloor+1}{(1-q)^{\nu}}, & \text{if } a<1\end{cases}\end{aligned} \tag{32}$$

with $a := (\nu\lfloor c-\rho\rfloor+q)/(1-\nu)$ and $q := c-\rho-\lfloor c-\rho\rfloor$. Indeed,

$$\inf_{\beta>0}\frac{\lfloor c-\rho+\beta\rfloor+1}{\beta^{\nu}} = \min_{i=1,2,\ldots}\frac{\lfloor c-\rho\rfloor+i}{(i-q)^{\nu}}. \tag{33}$$

with the mapping $g(x) := (\lfloor c-\rho\rfloor+x)/(x-q)^{\nu}$ being strictly decreasing in $(0,a)$ and strictly increasing in (a,∞), so that the mininum in (33) is reached when $\beta=\lfloor a\rfloor$ or when $\beta=\lceil a\rceil$ if $a\geq 1$ and when $\beta=1$ if $a<1$.

(iv) G is lognormal. The c.d.f. G of a r.v. σ is *lognormal* if $\sigma \stackrel{\text{st}}{=} \exp(Y)$ where Y is a Gaussian r.v. with mean μ and variance δ^2. Then, $\overline{G}(x) \sim (2\pi)^{-1/2}\,(\delta/(\log x-\mu))\; e^{-(\log x-\mu)^2/(2\,\delta^2)}$. From this we get

$$\overline{G_1}(x) \sim \frac{\delta^3\, x\, e^{-(\log x-\mu)^2/(2\,\delta^2)}}{\overline{\sigma}\,\sqrt{2\pi}\,(\log x-\mu)^2} \tag{34}$$

with $\overline{\sigma} = \exp(\mu+\delta^2/2)$. Proposition 3.2 yields a trivial lower bound (i.e. 0). From Proposition 3.1 (Corollary 3.1 does not apply since $G_1\notin\mathcal{D}$) we have

$$\liminf_{x\to\infty}\frac{1}{(\log x)^2}\,\log\, P(Q>x) \geq -\frac{\lfloor c-\rho\rfloor+1}{2\,\delta^2}. \tag{35}$$

4 Upper Bounds

We begin this section by stating two lemmas that will be used in the derivation of asymptotic upper bounds in the case when G and G_1 are subexponential probability distributions.

Lemma 4.1 (Cline [10]) *Let $F, F^1, \ldots, F^k$ be probability distributions such that $\overline{F^j}(x) \sim c_j \, \overline{F}(x)$, $c_j > 0$, for all $j = 1, 2, \ldots, k$. If $F \in \mathcal{S}$ then $\overline{F^1 \star \cdots \star F^k}(x) \sim \sum_{j=1}^{k} c_j \, \overline{F}(x)$.*

Lemma 4.2 (Pakes [26]) *Consider a GI/GI/1 queue with i.i.d. service times $\{\sigma_n\}_n$ with common c.d.f. F and i.i.d. interarrival times $\{\tau_n\}_n$. Assume that $E[\sigma_n] < E[\tau_n]$.*

If F, $F_1 \in \mathcal{S}$, then

$$P\left(W > x\right) \sim \frac{E[\sigma_n]}{E[\tau_n] - E[\sigma_n]} \, \overline{F_1}(x)$$

where $W := \sup_{n \in \mathrm{N}} \left(\sum_{m=0}^{n-1} \left(\sigma_m - \tau_m\right) \right)$ is the stationary waiting time.

We are now in position to derive the following asymptotic upper bounds for $P(Q > x)$ and for $\log P(Q > x)$ when G and G_1 are in $\mathcal{S}$.

Proposition 4.1 (Upper bounds)

Assume that $G, G_1 \in \mathcal{S}$. Then,

$$\limsup_{x \to \infty} \frac{P(Q > x)}{\overline{G_1}(x)} \leq \rho + \frac{\rho}{c - \rho}. \tag{36}$$

In particular, (36) implies that

$$\limsup_{x \to \infty} \frac{\log \, P(Q > x)}{-\log \overline{G_1}(x)} \leq -1. \tag{37}$$

Proof. Define

$$a_0 \;=\; \sum_{j=1}^{b^0} (\hat{\sigma}_j + 1) \tag{38}$$

$$a_s \;=\; \sum_{j=1}^{v_s} \lceil \sigma_j + T_j(s) - s \rceil, \quad s = 1, 2 \ldots \tag{39}$$

where v_s denotes the number of arrivals in the M/G/∞ queue in the interval of time $[s-1, s)$ and $T_j(s)$ is the time of the j-th arrival in $[s-1, s)$ for $s = 1, 2, \ldots$. Since the arrival process in this queue is Poisson with rate λ, $\{v_s,\, s \in \mathbb{N}\}$ constitutes an i.i.d. sequence of Poisson r.v's with intensity $E[\nu_s] = \lambda$, namely $P(a_s = k) = \lambda^k \exp(-k)/k!$ for all $k \in \mathbb{N}$.

We first establish some preliminary results related to the r.v.'s $a_0, a_1, \ldots$. To begin with, we observe from (8)-(9) and (38)-(39) that

(a1) $a_0(t) \leq a_0$ (a.s.) and $a_s(t) \leq_{st} a_s$ for all $t = 1, 2, \ldots$, $s = 1, 2, \ldots, t-1$;

(a2) the r.v.'s a_s, $s = 1, 2, \ldots$ are i.i.d. and independent of the r.v. a_0,

where $X \leq_{st} Y$ if the real-valued r.v.'s X and Y satisfy $E[f(X)] \leq E[f(Y)]$ for all measurable and nondecreasing mappings $f : (-\infty, \infty) \to (-\infty, \infty)$ such that the expectations exist.

To get the second inequality in (a1) note from (9) that

$$
\begin{aligned}
a_s(t) &= \sum_{s-1 \leq T_j < s} \sum_{i=0}^{t-1-s} I(\sigma_j + T_j - s > i) = \sum_{s-1 \leq T_j < s} \min(\lceil \sigma_j + T_j - s \rceil, t - s) \\
&\stackrel{d}{=} \sum_{j=1}^{v_s} \min(\lceil \sigma_j + T_j(s) - s \rceil, t - s) \leq \sum_{j=1}^{v_s} \lceil \sigma_j + T_j(s) - s \rceil = a_s \quad \text{for } s = 1, 2, \ldots, t-1,
\end{aligned}
$$

where $X \stackrel{d}{=} Y$ if the r.v.'s X and Y have the same probability distribution. Next, we focus on the asymptotic behavior of $P(a_s > x)$ for $s \in \mathbb{N}$. Under the assumptions $G, G_1 \in \mathcal{S}$, the inclusion $\mathcal{S} \subset \mathcal{L}$ (see Lemma 2.2(a)) and Lemma 2.2(c) imply that

$$
\begin{aligned}
\overline{G}(x) &= P(\sigma_j > x) \sim P(\sigma_j - 1 > x) \in \mathcal{S} & (40) \\
\overline{G_1}(x) &= P(\hat{\sigma}_j > x) \sim P(\hat{\sigma}_j + 1 > x) \in \mathcal{S}. & (41)
\end{aligned}
$$

On the other hand, the inequalities $\sigma_j - 1 \leq \lceil \sigma_j + T_j(s) - s \rceil \leq \sigma_j$ combined with (40) and Lemma 2.2(c) in turn yields

$$
\overline{G}(x) \sim P\left(\lceil \sigma_j + T_j(s) - s \rceil > x\right) \in \mathcal{S}. \tag{42}
$$

By using now (41), (42) and [14, Theorem 1.3.9] we see that

$$
\begin{aligned}
P(a_0 > x) &\sim \rho\, \overline{G_1}(x) & (43) \\
P(a_s > x) &\sim \lambda\, \overline{G}(x) \quad \text{for } s = 1, 2, \ldots. & (44)
\end{aligned}
$$

We conclude these preliminary remarks with the computation of $E[a_s]$ for $s \geq 1$. For fixed $s \geq 1$, the r.v. $s - T_j(s)$ is uniformly distributed over $(0, 1)$ (since the arrivals are Poisson) and independent of σ_j. Hence, by applying Lemma A.1 with $X = \sigma_j$ and $U = s - T_j$ we find that $E[\lceil \sigma_j + T_j - s \rceil] = E[\sigma_j]$, which in turn yields

$$
E[a_s] = E[\nu_s]\, E[\lceil \sigma_j + T_j - s \rceil] = \rho \tag{45}
$$

from Wald's identity and the definition of ρ.

We are now in position to proof (36). We start from (cf. (6), (10), (a1))

$$\begin{aligned} P(Q > x) &= P\left(\sup_{t\in\mathrm{N}}\left(a_0(t) + \sum_{s=1}^{t-1} a_s(t) - ct\right) > x\right) \\ &\le P\left(a_0 + \sup_{t\in\mathrm{N}}\left(\sum_{s=1}^{t} a_s - ct\right) > x\right) \\ &= P(a_0 + W > x) \end{aligned} \tag{46}$$

where a_0 and $W := \sup_{t\in\mathrm{N}}\left(\sum_{s=1}^{t} a_s - ct\right)$ are independent r.v.'s.

To proceed, we notice that under (a2), (45) and the (stability) condition $\rho < c$, $P(W \le x)$ is the probability distribution of the stationary waiting time in a stable $D/GI/1$ queue with interarrival times c and i.i.d. service times $\{a_s\}_s$. Therefore, by (44) and Lemma 4.2 [with $\sigma_n = a_n$ and $\tau_n = c$] we find

$$P(W > x) \sim \frac{\rho}{c - \rho}\overline{G_1}(x). \tag{47}$$

By using now (43), (46), (47), the independence of the r.v.'s a_0 and W (see (a2)), and Lemma 4.1 [with $F = G_1$, $F^1(x) = P(a_0 \le x)$ and $F^2(x) = P(W \le x)$] we conclude that (36) holds true.

∎

It is known that both G and G_1 belong to $\mathcal{S}$ when G is (i) Pareto, (ii) Weibull or (iii) lognormal. We conclude this section by specializing Proposition 4.1 to these particular probability distributions.

(i) G is Pareto. From (28) and (37) we get

$$\limsup_{x\to\infty} \frac{1}{\log x} \log P(Q > x) \le -\alpha + 1. \tag{48}$$

Also note that the bound in (48) is tighter than Duffield's corresponding bound (3) when $c - \rho \le \alpha/(\alpha - 1)$; otherwise Duffield's is tighter.

(ii) G is Weibull. From (31) and (37) we get

$$\limsup_{x\to\infty} \frac{1}{x^\nu} \log P(Q > x) \le -1. \tag{49}$$

(iii) G is lognormal. From (34) and (37) we get

$$\limsup_{x\to\infty} \frac{1}{(\log x)^2} \log P(Q > x) \le -\frac{1}{2\delta^2}. \tag{50}$$

We observe from (29), (48) and (35), (50) that the bounds are tight when $c - \rho < 1$:

Corollary 4.1 *Assume that $c - \rho < 1$. If G is Pareto then*

$$\lim_{x\to\infty} \frac{1}{\log x} \log P(Q > x) = -\alpha + 1 \tag{51}$$

and if G is lognormal then

$$\lim_{x\to\infty} \frac{1}{(\log x)^2} \log P(Q > x) = -\frac{1}{2\delta^2}. \tag{52}$$

5 Concluding Remarks

We conclude this paper by addressing the situation when the multiplexer is fed by N independent M/G/∞ input processes, with arrival rate λ_i and c.d.f. of the service times G^i for the system i ($i = 1, 2, \dots, N$). Because the arrivals are Poisson this is equivalent to considering a single M/G/∞ queueing system with arrival intensity $\lambda := \sum_{i=1}^N \lambda_i$ and c.d.f. G of the service time given by $G(x) = \sum_{i=1}^N (\lambda_i/\lambda)\, G^i(x)$. All of the results in the paper therefore apply to this pair (λ, G). Of particular interest is the case when one c.d.f. of the service times, say G^1, has a heavier tail than the others, namely, $\overline{G^i}(x) = o(\overline{G^1}(x))$ for all $i = 2, 3, \dots, N$. Then, $\overline{G_1}(x) \sim (\lambda_1/\lambda)\, \overline{G_1^1}(x)$ and we may conclude from the results in Sections 3-4 that the source with the heaviest tail dominates the other sources. In particular, we see from (11) and (37) that

$$-\theta_1 \le \liminf_{x\to\infty} \frac{\log P(Q > x)}{-\log \overline{G_1^1}(x)} \le \limsup_{x\to\infty} \frac{\log P(Q > x)}{-\log \overline{G_1^1}(x)} \le -1$$

where the upper bound holds if $G^1, G_1^1 \in \mathcal{S}$, with $\theta_1 := \inf_{\beta>0} \left\{ h(\beta) \limsup_{x\to\infty} \frac{\log \overline{G_1^1}(x)}{\log \overline{G_1^1}(\beta x)} \right\}$, $h(\beta) := \lfloor c - \rho + \beta \rfloor + 1$ and $\rho = \sum_{i=1}^N (\lambda_i/\lambda) \int_0^\infty x\, G^i(dx)$.

Acknowledgements: The authors would like to thank Rajeev Agrawal for a useful discussion during the course of this work.

A Appendix

Lemma A.1 *Let X and U be independent r.v.'s. We assume that U is uniformly distributed over $(0, 1)$ and X is a nonnegative r.v. Then,*

$$E\left[\lceil X - U \rceil\right] = E[X]. \tag{53}$$

Proof. Since $\lceil X - U \rceil$ is a nonnegative integer, we have

$$E\left[\lceil X - U \rceil\right] = \sum_{n\ge 0} P(\lceil X - U \rceil > n)$$

$$
\begin{aligned}
&= \sum_{n\geq 0} P(X-U>n) = \sum_{n\geq 0} \int_0^1 P(X>n+u)\,du \\
&= \sum_{n\geq 0} \int_n^{n+1} P(X>u)\,du = \int_0^\infty P(X>u)\,du = E[X].
\end{aligned}
$$

■

References

[1] V. Anantharam, "On the sojourn time of sessions at an ATM buffer with long-range dependent input traffic," *Proc. of the 34th IEEE Conf. on Decision & Control*, New Orleans, LA, Dec. 13-15, 1995, Vol. 1, 859-864.

[2] J. Beran, *Statistics for long-memory processes.* Chapman and Hall, New York, 1994.

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

[4] A. A. Borovkov, *Stochastic processes in queueing theory.* Springer-Verlag, New York, 1976.

[5] O. J. Boxma, "Fluid queues and regular variation," *Performance Evaluation*, **27**&**28**:699-712, 1996.

[6] O. J. Boxma, V. Dumas, "Fluid queues with long-tailed activity period distributions." Preprint 1997. To appear in a Special Issue of *Computer Communications* on "Stochastic Analysis and Optimisation of Communication Systems".

[7] F. Brichet, J. W. Roberts, A. Simonian, D. Veitch, "Heavy traffic analysis of a storage model with long range dependent on/off sources", Preprint 1996. To appear in *QUESTA*.

[8] V. P. Chistakov, "A theorem on sums of independent positive random variables and its application to branching random processes," *Theory Prob. Appl.*, **9**:640-648, 1964.

[9] G. L. Choudhury, W. Whitt. "Long-tail buffer-content distributions in broadband networks". To appear in *Performance Evaluation.*

[10] D. B. H. Cline, "Convolution tails, product tails and domains of attraction," *Prob. Theory Related Fields*, **72**:529-557, 1986.

[11] D. R. Cox, V. Isham, *Point processes.* Chapman and Hall, New York, 1980.

[12] N. G. Duffield, "On the relevance of long-tailed durations for the statistical multiplexing of large aggregations," *Proc. of the 34th Annual Allerton Conf. on Communication, Control and Computing*, Oct. 2-4, 1996.

[13] N. G. Duffield, N. O'Connell, "Large deviations and overflow probabilities for the general single-server queue with applications," *Math. Proc. Cam. Phil. Soc.*, **118**:363-374, 1995.

[14] P. Embrechts, C. Klüppelberg, T. Mikosch, *Modelling Extremal Events.* Springer-Verlag, Berlin, 1997.

[15] P. Embrechts, E. Omey, "A property of longtailed distributions," *J. Appl. Prob.*, **21**:80-87, 1984.

[16] M. Garrett, W. Willinger, "Analysis, modeling and generation of self- similar VBR video traffic," *Proc. SIGCOMM'94*, 269-280, Sep. 1994.

[17] C. M. Goldie, "Subexponential distributions and dominated-variation tails," *J. Appl. Prob.*, **15**:440-442, 1978.

[18] D. Heath, S. Resnick and G. Samorodnitsky, "Patterns of buffer overflow in a class of queues with long memory in the input stream." Preprint 1996.

[19] P. R. Jelenkovic, A. A. Lazar, "Asymptotic results for multiplexing subexponential on-off sources." Submitted *Adv. in Appl. Prob.*.

[20] F. P. Kelly, *Reversibility and Stochastic Networks.* John Wiley & Sons, New York, 1979.

[21] W. Leland, M. Taqqu, W. Willinger, D. Wilson, "On the self-similar nature of ethernet traffic(extended version)," *IEEE/ACM Trans. on Networking* **2**:1-15, 1994.

[22] N. Likhanov, B. Tsybakov, N. D. Georganas, "Analysis of an ATM buffer with self-similar ("fractal") input traffic", *Proc. of the IEEE Infocom'95 Conf.*, Boston, MA, Apr. 4-6, 1995, 985-992.

[23] Z. Liu, P. Nain, D. Towsley and Z.-L. Zhang, "Asymptotic behavior of a multiplexer fed by a long-range dependent process". [http://www.inria.fr/mistral/personnel/Philippe.Nain/moi.html/research.html]

[24] J. Neveu, *Bases Mathématiques du Calcul des Probabilités.* Masson & Cie, Paris, 1979.

[25] I. Norros, "A storage model with self-similar input," *QUESTA*, **16**:387-396, 1994.

[26] A. G. Pakes, "On the tails of waiting time distributions," *J. Appl. Prob.*, **12**:555-564, 1975.

[27] M. Parulekar, A. M. Makowski, "Tail probabilities for a multiplexer with self-similar traffic," *Proc. of the IEEE Infocom'96 Conf.*, San Francisco, CA, Mar. 26-28, 1996, 1452-1459.

[28] M. Parulekar, A. M. Makowski, "Tail probabilities for $M/G/\infty$ input processes (I): Preliminary asymptotics." To appear in *QUESTA*.

[29] M. Parulekar, A. M. Makowski, "$M/G/\infty$ input processes: A versatile class of models for network traffic," *Proc. of the IEEE Infocom'97 Conf.*, Kobe, Japan.

[30] V. Paxson, S. Floyd, "Wide area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. on Networking*, **3**:226-244, 1993.

[31] T. Rolski, S. Schelgel and V. Schmidt, "Asymptotics of Palm-stationary buffer content distributions in fluid flow queues." Preprint Nov. 1996.

[32] A. N. Shiryayev, *Probability.* Springer-Verlag, New York, 1984.

[33] L. Takács, *Theory of queues.* Oxford University Press, New York, 1962.