Technical Research Report

Buffer Overflow Probabilities for a Multiplexer with Self-Similar Traffic

by M. Parulekar, A.M. Makowski

CSHCN T.R. 95-25
(ISR T.R. 95-67)

Buffer overflow probabilities for a multiplexer with self–similar traffic

MINOTHI PARULEKAR [†]
minothi@eng.umd.edu
(301) 405-2948

ARMAND M. MAKOWSKI [‡]
armand@eng.umd.edu
(301) 405-6648
FAX:(301) 314-9281

Abstract

We study the large buffer asymptotics of a multiplexer under two different self–similar traffic inputs, namely the so–called $M|G|\infty$ model of Cox and the fractional Gaussian noise input model. In the former case we show that the tail probabilities for the buffer content, (in steady–state) decay at most hyperbolically. This is to be contrasted with the situation where the input traffic is fractional Gaussian noise, in which case the tail probabilities display a Weibullian character. Therefore, for a given input rate r_{in} and Hurst parameter H, these dissimilar asymptotics would result in vastly differing buffer engineering practices, which points somewhat to the inadequacy of using H as the sole parameter to characterize long–range dependence.

1 Introduction

In recent years increasing evidence has accumulated that points to the (asymptotically) self–similar nature of aggregate packet streams in a wide range of currently working packet networks, e.g., Ethernet LANs [9, 14, 22], VBR traffic [3], WAN traffic [10, 20]. This self–similarity manifests itself most crisply through a long–range

[†] Electrical Engineering Department and Institute for Systems Research, University of Maryland, College Park, MD 20742. The work of this author was supported through NSF Grant NSFD CDR-88-03012.

[‡] Electrical Engineering Department and Institute for Systems Research, University of Maryland, College Park, MD 20742. The work of this author was supported partially through NSF Grant NSFD CDR-88-03012 and through NASA Grant NAGW277S.

dependence effect [2, 4] which is characterized by the autocorrelation of the traffic process obeying a power law (in the lag time). Long–range dependent processes are inherently non–Markovian, and have the property that while long–term correlations are individually small, they nevertheless accumulate in the long run to create scenarios which are drastically different from those produced by more traditional, typically Markovian in nature, short–range dependent models.

This established presence of long–range dependence over a wide range of time scales in packet traffic processes is expected to have an impact on queuing performance and traffic engineering. In this paper, we seek to gain some insights into these issue by investigating the effects of long–range dependence on the behavior of the buffer queue at a multiplexer. More precisely, we consider a discrete–time single server queue with infinite capacity and constant release rate of c cells/slot, as a surrogate for a multiplexer, and feed it with a (stationary) traffic stream which exhibits some (possibly asymptotic) self–similarity; two traffic models are investigated, namely the so–called $M|G|\infty$ model of Cox and the fractional Gaussian noise input model.

In particular, if q_∞ denotes the steady–state buffer content (in number of cells) at the multiplexer, we are interested in tail probabilities $\mathbf{P}\left[q_\infty > b\right]$ for large b as a means to estimating buffer overflow probabilities for the corresponding finite buffer system. Such asymptotics are often the first guiding step to size up the buffer at the multiplexer in order to guarantee quality of service requirements. We briefly review our findings under the two sets of traffic assumptions mentioned above.

The $M|G|\infty$ model appeared to have been mentioned first by Cox in [4] as an example of an asymptotically self–similar process. It is obtained by generating customers according to a (discrete–time) "Poisson" process and by offering them to an infinite server group under the assumption that the common distribution of service durations has a heavy tail, say a discrete Pareto distribution with parameter α, $1 < \alpha < 2$. The process that counts the number of busy servers at the beginning of a time slot is what we refer to as the $M|G|\infty$ input model. This process, or rather its stationary version, is an asymptotically self–similar process with Hurst parameter $H = (3-\alpha)/2$. With the $M|G|\infty$ traffic feeding into the multiplexer, we show that the $M|G|\infty$ input induces an asymptotic regime at the buffer in steady state which is characterized by

$$\mathbf{P}\left[q_\infty > b\right] \geq b^{-(1-2H)(c-r_{in})+o(1)} \quad (b \to \infty) \tag{1.1}$$

where r_{in} denotes the average input rate, i.e., the average number of cells offered per slot. The derivation of (1.1) is an application of some recent results of Duffield and O'Connell [7] on the buffer asymptotics of a general single–server queue. Although we expect (1.1) to hold as an equality, we were unable to establish this fact as of the writing of this paper. In any case, these asymptotics already indicate that the tail probabilities $\mathbf{P}[q_\infty > b]$ display a (negative) power law with exponent $\alpha - 1$, in sharp contrast with the geometric decay that is usually observed under Markovian, thus short–range dependent, input streams. A similar result was reported in [15] where an aggregate traffic model was constructed by superposing a large number of on–off sources with Pareto distributed activity periods. In Section 6 we argue that the limiting model of [15] is nothing else but the $M|G|\infty$ model of Cox. However, the asymptotics of [15] are for a somewhat different model as will be explained in Section 6.

On the other hand, the fractional Gaussian noise input model, which constitutes the discrete–time analog of fractional Brownian motion, is perhaps the simplest of self–similar processes to describe, as it is a zero–mean Gaussian process with stationary increments and covariance structure (7.1). If the input stream is modeled as a fractional Gaussian noise process with Hurst parameter H, $0 < H < 1$, and with "drift" r_{in}, then it can be shown that

$$\mathbf{P}[q_\infty > b] = \exp\left(-\frac{1}{2\sigma^2} \left(\frac{(c - r_{in})^H}{C_H} \right)^2 b^{2(1-H)} (1 + o(1)) \right) \quad (b \to \infty) \qquad (1.2)$$

This time the asymptotics of the steady–state buffer follow a Weibull–like distribution, as was announced [3]. A qualitatively similar result was obtained by Norros [17] for a continuous–time storage model driven by fractional Brownian motion. The asymptotics there, as well as (1.2), follow from the results of Duffield and O'Connell [7].

As has been reported by several authors [3, 8], it is already apparent from (1.1)–(1.2) that long–range dependence in the input traffic will indeed induce buffer dynamics which are qualitatively very different from those which arise in Markovian models. However, even a cursory comparison of (1.1) and (1.2) suggests a much richer range of possibilities even within the class of (asymptotically) self–similar input models. In fact, for given input rate r_{in} and Hurst parameter H, these very dissimilar asymptotics would result in vastly differing buffer engineering practices, and this points somewhat to the inadequacy of using the Hurst parameter as the sole parameter to characterize long–range dependence.

Of course, it could be argued at this point that some (asymptotically) self–similar input models are more appropriate than others. However, as a quick review of the existing literature indicate, both fractional Gaussian noise and $M|G|\infty$ input models have provided good fits for diverse applications. The fractional Gaussian noise input model and its continuous–time analog are convenient mathematically [1, 7, 17], yet are alleged [8] to provide a reasonably good statistical fit to actual measurements in data networks. On the other hand, the $M|G|\infty$ input model has been found to match reasonably well some wide area applications [20].

The paper is organized as follows: The queuing model and various preliminaries are presented in Section 2. Existing results on buffer asymptotics are summarized in Section 3 for easy reference. We introduce the $M|G|\infty$ input model in Section 4, and the corresponding buffer asymptotics are developed in Section 5; the calculations, which are fairly involved, are outlined in Section 8. In Section 6 we compare the material of Sections 4 and 5 with some recent results on aggregate models of on–off sources. Finally, in Section 7 we obtain the buffer asymptotics when the input is modeled by a fractional Gaussian noise process.

A few words on the notation used in this paper: We denote the set of non–negative integers by $I\!N$, and the set of all real (resp. non–negative real) numbers by $\mathbb{R}$ (resp. $\mathbb{R}_+$). For any scalar x in $\mathbb{R}$, we write $[x]$ to denote the integer part or floor of x. All rvs are defined on some probability triple $(\Omega, \mathcal{F}, \mathbf{P})$, with $\mathbf{E}$ denoting the corresponding expectation operator. Finally two rvs X and Y are said to be *equal in law* if they have the same distribution, a fact we denote by $X =_{st} Y$. Weak convergence is denoted by $\Longrightarrow$.

2 The model and preliminaries

We map the multiplexer into a discrete–time single server queue with infinite buffer capacity which operates at a constant rate and in a first–come first–served manner: Let q_t denote the number of cells remaining in the buffer by the end of slot $[t-1, t)$, and let b_{t+1} denote the number of new cells which arrive at the start of time slot $[t, t+1)$. If the multiplexer output link can transmit c cells/slot, then the buffer content sequence $\{q_t, \ t = 0, 1, \ldots\}$ evolves according to Lindley recursion

$$q_0 = 0; \quad q_{t+1} = [q_t + b_{t+1} - c]^+, \quad t = 0, 1, \ldots \tag{2.1}$$

This system admits a steady–state regime under the following very broad conditions which are due to Loynes [16].

Proposition 2.1 *Assume the $\mathbb{N}$–valued arrival sequence $\{b_{t+1},\ t = 0, 1, \ldots\}$ to be stationary and ergodic. If $\mathbf{E}[b_1] < c$, then the system is stable in the sense that $q_t \Longrightarrow_t q_\infty$ for some $\mathbb{R}_+$–valued rv q_∞.*

Throughout, when considering the model (2.1), we shall at a minimum, assume that the conditions of Proposition 2.1 are satisfied. We shall often write $r_{in} \equiv \mathbf{E}[b_1]$ to stress the fact that $\mathbf{E}[b_1]$ indeed represents the average input rate into the system. This will come handy when comparing the effect on the multiplexer of several traffic streams with a given value for r_{in} such that $r_{in} < c$.

Recently, there has been considerable interest in estimating the tail probabilities $\mathbf{P}[q_\infty > b]$ for large b, as a means of estimating buffer overflow probabilities for the corresponding finite buffer system. In the next section we summarize some of these results as they apply to our setup. These asymptotics make use of the representation

$$q_\infty =_{st} \sup\{S_t - ct,\ t = 0, 1, \ldots\}. \qquad (2.2)$$

with

$$S_0 = 0; \quad S_t = b_1 + \ldots + b_t, \quad t = 1, 2, \ldots \qquad (2.3)$$

Taking this representation for q_∞ as a point of departure, several authors have obtained estimates on the tail probabilities by means of large deviations estimates for the sequence $\{S_t - ct,\ t = 0, 1, \ldots\}$.

To fix the terminology used in this paper, we recall the notion of a Large Deviations Principle as discussed in [6]: Consider a monotone increasing $\mathbb{R}$–valued sequence $\{v_t,\ t = 0, 1, \ldots\}$ such that $\lim_{t\to\infty} v_t = \infty$. A sequence of $\mathbb{R}$–valued rvs $\{x_t,\ t = 0, 1, \ldots\}$ is said to satisfy the *Large Deviations Principle under scaling* v_t if there exists a lower–semicontinuous function $I : \mathbb{R} \to [0, \infty]$ such that for every open set G,

$$-\inf_{x\in G} I(x) \le \liminf_{t\to\infty} \frac{1}{v_t} \ln \mathbf{P}[x_t \in G] \qquad (2.4)$$

and for every closed set F,

$$\limsup_{t\to\infty} \frac{1}{v_t} \ln \mathbf{P}[x_t \in G] \le -\inf_{x\in F} I(x). \qquad (2.5)$$

Usually it is required that the rate function I be level compact in the sense that the level sets of I all be compact, i.e., for each $r > 0$, the set $\{x \in \mathbb{R} :\ I(x) \le r\}$ is a compact subset of $\mathbb{R}$. In that case, the rate function I is said to be good.

In many situations of interest, the rate function can be expressed as the Legendre–Fenchel transform $\Lambda^\star$ of another mapping $\Lambda : \mathbb{R} \to (-\infty, \infty]$, namely

$$\Lambda^\star(z) \equiv \sup_{\theta \in \mathbb{R}} \{\theta z - \Lambda(\theta)\}, \quad z \in \mathbb{R}. \tag{2.6}$$

The reader is referred to the monograph [6] for additional information on the subject matter of Large Deviations.

3 Buffer asymptotics

We begin by summarizing recent results on tail probabilities which have been obtained by several authors in varying degrees of generality [11], [13]. These results pave the way for the definition of the notion of effective bandwidth [12], [21].

We introduce

$$\Lambda_t(\theta) \equiv \frac{1}{t} \ln \mathbf{E}\left[\exp(\theta(S_t - ct))\right], \quad \theta \in \mathbb{R} \tag{3.1}$$

for each $t = 1, 2, \ldots$; by Jensen's inequality it is plain that $\Lambda_t(\theta) \geq \theta \mathbf{E}[b_1]$.

Proposition 3.1 *Assume the $\mathbb{N}$–valued arrival sequence $\{b_{t+1},\ t = 0, 1, \ldots\}$ to be stationary and ergodic, and to satisfy the following conditions:*

1. *The limit*

$$\Lambda(\theta) \equiv \lim_{t \to \infty} \Lambda_t(\theta), \quad \theta \in \mathbb{R} \tag{3.2}$$

exists (possibly as an extended real number);

2. *The set $\Theta \equiv \{\theta > 0 : \Lambda(\theta) < 0\}$ is non–empty, and*

$$\Lambda_t(\theta) < \infty, \quad \theta \in \Theta,\ t = 1, 2, \ldots \tag{3.3}$$

3. *The process $\{t^{-1}(S_t - ct),\ t = 1, 2, \ldots\}$ satisfies the Large Deviations Principle under scaling t with good rate function $\Lambda^\star$.*

Then

$$\lim_{b \to \infty} \frac{1}{b} \ln \mathbf{P}\left[q_\infty > b\right] = -\theta^\star \tag{3.4}$$

where

$$\theta^\star = \sup\{\theta > 0 : \Lambda(\theta) < 0\} = \inf_{y > 0} \frac{\Lambda^\star(y)}{y}. \tag{3.5}$$

It is helpful to reformulate these facts by introducing the auxiliary quantities

$$\Lambda^b_t(\theta) \equiv \frac{1}{t} \ln \mathbf{E} \left[\exp(\theta S_t) \right] \quad \theta \in \mathbb{R} \tag{3.6}$$

for each $t = 1, 2, \ldots$, whence $\Lambda_t(\theta) = \Lambda^b_t(\theta) - c\theta$. Under the assumptions of Proposition 3.1, we find that the limit $\Lambda_b(\theta) \equiv \lim_{t\to\infty} \Lambda^b_t(\theta) = \Lambda(\theta) + c\theta$ exists (possibly as an extended real number); the set Θ can be described by $\Theta = \{\theta > 0 : \Lambda_b(\theta) < c\theta\}$. Finally, the process $\{t^{-1}S_t,\ t = 1, 2, \ldots\}$ satisfies the Large Deviations Principle with good rate function $\Lambda^\star_b$, and the conclusion (3.5) can be expressed as

$$\theta^\star = \sup\{\theta > 0 : \Lambda_b(\theta) < c\theta\} = \inf_{y>0} \frac{\Lambda^\star_b(c+y)}{y}. \tag{3.7}$$

When applied to input processes with long-range dependence, Proposition 3.1 fails and needs to be modified accordingly. Duffield and O'Connell [7] have recently generalized this result by allowing two different scalings, one for the law of large numbers associated with S_t and one for exponential decay of rare events. We begin with two strictly monotone increasing $\mathbb{R}_+$-valued sequences $\{v_t,\ t = 1, 2, \ldots\}$ and $\{a_t,\ t = 1, 2, \ldots\}$ such that $\lim_t v_t = \lim_t a_t = \infty$, and modify (3.1) to read

$$\Lambda_t(\theta) \equiv \frac{1}{v_t} \ln \mathbf{E} \left[\exp\left(\theta v_t (\frac{S_t - ct}{a_t}) \right) \right], \quad \theta \in \mathbb{R}. \tag{3.8}$$

The inverse of $\{a_t,\ t = 1, 2, \ldots\}$ is the mapping $a^{-1} : \mathbb{R}_+ \to I\!N$ defined by $a^{-1}(x) \equiv \sup\{k \in I\!N : a_k \le x\}$ for all $x \ge 0$. We also assume that there exist functions $g, h : \mathbb{R}_+ \to \mathbb{R}_+$ such that h is monotone increasing with $\lim_{b\to\infty} h(b) = \infty$ and the limit

$$\lim_{b\to\infty} \frac{v_{a^{-1}(b/y)}}{h(b)} = g(y), \quad y > 0 \tag{3.9}$$

holds. The following is essentially Theorem 2.1 obtained by Duffield and O'Connell in [7].

Proposition 3.2 *Assume the arrival sequence $\{b_{t+1},\ t = 0, 1, \ldots\}$ to be stationary and ergodic, and to satisfy the following conditions:*

1. *The limit*

$$\Lambda(\theta) \equiv \lim_{t\to\infty} \Lambda_t(\theta), \quad \theta \in \mathbb{R} \tag{3.10}$$

exists;

2. *The process $\{t^{-1}(S_t - ct),\ t = 1, 2, \ldots\}$ satisfies the Large Deviations Principle with good rate function $\Lambda^\star$ under scaling v_t.*

Then, for each $y > 0$ we have

$$\liminf_{b\to\infty} \frac{1}{h(b)} \ln \mathbf{P}\left[q_\infty > b\right] \geq -g(y) \inf_{x>y} \Lambda^\star(x) \tag{3.11}$$

In general, the rate function $\Lambda^\star$ is only lower–semicontinuous so that

$$\inf_{x>y} \Lambda^\star(x) = \Lambda^\star(y+), \quad y > 0 \tag{3.12}$$

with $\Lambda^\star(y+)$ denoting the left limit of $\Lambda^\star$ at y; such a left limit exists by the convexity of $\Lambda^\star$. In particular, if the rate function $\Lambda^\star$ is continuous on $[0,\infty)$, then Proposition 3.2 immediately implies the lower bound

$$\liminf_{b\to\infty} \frac{1}{h(b)} \ln \mathbf{P}\left[q_\infty > b\right] \geq -\gamma^\star \tag{3.13}$$

with $\gamma^\star$ given by

$$\gamma^\star \equiv \inf_{y>0} g(y)\Lambda^\star(y). \tag{3.14}$$

In [7], under additional conditions to the ones of Proposition 3.2, the authors derive a companion upper bound to (3.11)–(3.13). Although, the lower and upper bounds may in principle not be tight, in many situations they will be, thereby giving asymptotics of the form

$$\liminf_{b\to\infty} \frac{1}{h(b)} \ln \mathbf{P}\left[q_\infty > b\right] = -\gamma^\star \tag{3.15}$$

for some positive constant $\gamma^\star$; this constant is usually given by (3.14).

Unfortunately, for one of the cases of interest here, the one described in Section 4, the conditions, as given in [7], under which the upper bound holds, are not satisfied.

4 The $M|GI|\infty$ input model

In order to model long–range dependence, we make use of a model which appears to have first been described by Cox [4]: Consider the following discrete–time infinite server set–up. During time slot $[t, t+1)$, β_{t+1} new customers arrive into the system. Customer i, $i = 1, \ldots, \beta_{t+1}$, is presented to its own server and begins service by the start of slot $[t+1, t+2)$; its service time has duration $\sigma_{t+1,i}$. Let b_t denote the number of busy servers, or equivalently of customers still present in the system, at the beginning of slot $[t, t+1)$, with b denoting the number of busy servers initially present in the system at $t = 0$. Under conditions which are now specified, the process $\{b_t,\ t = 0, 1, \ldots\}$ exhibits long–range dependence.

Consider the $I\!N$-valued rvs r, $\{\beta_{t+1},\ t = 0,1,\ldots\}$ and $\{\sigma_{t,i},\ t = 0,1,\ldots;\ i = 1,2,\ldots\}$ under the following assumptions: (i) The rvs are mutually independent; (ii) The rvs $\{\beta_{t+1},\ t = 0,1,\ldots\}$ are *i.i.d.* Poisson rvs with parameter $\lambda > 0$; (iii) The rvs $\{\sigma,\ \sigma_{t,i},\ t = 0,1,\ldots;\ i = 1,2,\ldots\}$ are *i.i.d.* with common pmf G on $\{1,2,\ldots\}$ – this pmf G will be shortly specified.

The first indication that the rvs $\{b_t,\ t = 0,1,\ldots\}$ exhibit some form of dependence can already be traced to the fact that these rvs are indeed positively correlated in a strong sense: For all $t = 0,1,\ldots$, we write $b^t \equiv (b_0, b_1, \ldots, b_t)$.

Proposition 4.1 *The rvs $\{b_t,\ t = 0,1,\ldots\}$ are associated, in that for any $t = 0,1,\ldots$ and any pair of non-decreasing mappings $f,g : I\!N^{t+1} \to \mathbb{R}$,*

$$\mathbf{E}\left[f(b^t)g(b^t)\right] \geq \mathbf{E}\left[f(b^t)\right]\mathbf{E}\left[g(b^t)\right] \tag{4.1}$$

provided the expectations exist and are finite.

In particular, this fact, which is established in [18], [19], already implies

$$\Gamma(t,s) \equiv \text{cov}\left[b_t, b_s\right] \geq 0, \quad s,t = 0,1,\ldots \tag{4.2}$$

In what follows we take the pmf $G = \{g_r,\ r = 1,2,\ldots\}$ to be a Pareto distribution with parameter α, $1 < \alpha < 2$, in the sense that

$$\lim_{r\to\infty} \frac{\mathbf{P}\left[\sigma > r\right]}{r^{-\alpha}} = 1. \tag{4.3}$$

In this paper, for sake of concreteness we shall use

$$g_r \equiv \mathbf{P}\left[\sigma = r\right] = r^{-\alpha} - (r+1)^{-\alpha}, \quad r = 1,2,\ldots \tag{4.4}$$

in which case

$$\mathbf{P}\left[\sigma > r\right] = (r+1)^{-\alpha}, \quad r = 0,1,\ldots \tag{4.5}$$

Simple calculations then show that

$$\mathbf{E}\left[\sigma\right] = \sum_{r=0}^{\infty} \mathbf{P}\left[\sigma > r\right] = \sum_{r=1}^{\infty} r^{-\alpha} < \infty. \tag{4.6}$$

while $\mathbf{E}\left[\sigma^2\right] = \infty$. The following facts can be shown [4], [5]:

Proposition 4.2 *If b is a Poisson rv with parameter $\lambda\mathbf{E}\left[\sigma\right]$, then the rvs $\{b_t,\ t = 0,1,\ldots\}$ form a stationary sequence with*

$$\Gamma(h) \equiv \Gamma(t,t+h) = \lambda\mathbf{E}\left[(\sigma - h)^+\right], \quad t,h = 0,1,\ldots \tag{4.7}$$

Hence,

$$\Gamma(h) = \lambda \sum_{n=h+1} n^{-\alpha}, \quad h = 0, 1, \ldots \tag{4.8}$$

and standard bounding arguments yield

$$\lim_{h \to \infty} \frac{\Gamma(h)}{h^{\alpha-1}} = \lambda(\alpha - 1)^{-1}. \tag{4.9}$$

Therefore, in the notation of [4], this stationary sequence $\{b_t,\ t = 0, 1, \ldots\}$ is asymptotically self–similar with Hurst parameter H given by

$$H = 1 - \frac{1}{2}(\alpha - 1) = \frac{1}{2}(3 - \alpha). \tag{4.10}$$

Note that $0.5 < H < 1$ if $1 < \alpha < 2$, whence the process $\{b_t,\ t = 0, 1, \ldots\}$ exhibits long–range dependence [2], [4].

Under the assumptions of Proposition 4.2, each of the rvs $\{b_t,\ t = 0, 1, \ldots\}$ is a Poisson rv with parameter $\lambda\mathbf{E}[\sigma]$, whence

$$r_{in} \equiv \mathbf{E}[b_t] = \lambda\mathbf{E}[\sigma], \quad t = 0, 1, \ldots \tag{4.11}$$

5 Asymptotics for the $M|GI|\infty$ input model

We shall now apply Proposition 3.2 to the situation when the recursion is driven by the $M|G|\infty$ model introduced in the previous section. Here, the appropriate scaling is provided by $v_t = \ln t$ and $a_t = t$ for all $t = 1, 2, \ldots$, so that $a^{-1}(x) = [x]$ for all $x \geq 0$. Therefore, $v_{a^{-1}(b/y)} = \ln[b/y]$ and the choice $h(b) = \ln b$ yields $g(y) \equiv 1$ for all $y > 0$. In [18], [19], we show the following result: Set

$$\Lambda_t^b(\theta) \equiv \frac{1}{v_t} \ln \mathbf{E}\left[\exp(\theta v_t \frac{S_t}{t})\right], \quad \theta \in \mathbb{R} \tag{5.1}$$

for all $t = 1, 2, \ldots$, and note that,

$$\Lambda_t(\theta) = \Lambda_t^b(\theta) - c\theta, \quad \theta \in \mathbb{R}. \tag{5.2}$$

Proposition 5.1 *Under the foregoing assumptions, the limit*

$$\Lambda_b(\theta) = \lim_{t \to \infty} \Lambda_t^b(\theta), \quad \theta \in \mathbb{R} \tag{5.3}$$

exists and is given by

$$\Lambda_b(\theta) = \begin{cases} \lambda\mathbf{E}[\sigma]\,\theta & \text{if}\ \theta \leq \alpha - 1 \\ \infty & \text{if}\ \theta > \alpha - 1. \end{cases} \tag{5.4}$$

An outline of the calculations leading to (5.4) is presented in Section 8. On observing from (5.2) that

$$\begin{aligned}\Lambda(\theta) &= \Lambda_b(\theta) - c\theta \\ &= \begin{cases} (\lambda \mathbf{E}[\sigma] - c)\,\theta & \text{if } \theta \le \alpha - 1 \\ \infty & \text{if } \theta > \alpha - 1, \end{cases}\end{aligned} \tag{5.5}$$

we readily obtain that

$$\begin{aligned}\Lambda^\star(z) &= \sup_{\theta \in \mathbb{R}} \{\theta z - \Lambda(\theta)\} \\ &= \sup_{\theta \le \alpha - 1} \{\theta z - (\lambda \mathbf{E}[\sigma] - c)\theta\} \\ &= (\alpha - 1)\,(z + c - \lambda \mathbf{E}[\sigma])^+, \quad z \in \mathbb{R}.\end{aligned} \tag{5.6}$$

By the Gärtner–Ellis Theorem [6] the process $\{t^{-1}(S_t - ct),\ t = 1, 2, \ldots\}$ does satisfy the Large Deviations Principle with good rate function $\Lambda^\star$ under scaling $\ln t$, and the assumptions of Proposition 3.2 both hold.

Consequently, the rate function $\Lambda^\star$ being continuous on $\mathbb{R}$, from the remarks around (3.14) we conclude that

$$\begin{aligned}\gamma^\star = \inf_{z>0} g(z)\Lambda^\star(z) &= \inf_{z>0} (\alpha - 1)\,(z + c - \lambda \mathbf{E}[\sigma])^+ \\ &= (\alpha - 1)\,(c - r_{in})^+ \\ &= (\alpha - 1)\,(c - r_{in}) > 0\end{aligned} \tag{5.7}$$

where in the last step we have used the stability condition $r_{in} = \lambda \mathbf{E}[\sigma] < c$. Hence, (3.11) becomes

$$\liminf_{b \to \infty} \frac{1}{\ln b} \ln \mathbf{P}[q_\infty > b] \ge -(\alpha - 1)\,(c - r_{in}). \tag{5.8}$$

In other words, the buffer overflow probability satisfies

$$\mathbf{P}[q_\infty > b] \ge b^{-(\alpha-1)(c-r_{in})+o(1)} \quad (b \to \infty) \tag{5.9}$$

and the decay is therefore at best hyperbolic; this was announced in [15]. Even in the absence of the complementary upper bound, the calculations given here for the $M|GI|\infty$ input model already suggest that the asymptotic tail probabilities are not Weibull–like as was suggested in [3].

6 Comparison with an aggregate model of on–off sources

In a recent paper [15] an aggregate traffic model was constructed by superposing a large number of on–off sources with Pareto distributed activity periods. More precisely, consider M i.i.d. on–off sources such that during any time slot, source i, $i = 1, \ldots, M$, can be in one of two states, i.e., active or idle. In an active period, the source generates cells at rate R (bits/slot), while in the idle state the source is quiescent and does not generate cells. Expressed in numbers of slots, the lengths of the ℓ^{th} active and idle periods are denoted by $\tau_\ell^{(i)}$ and $\theta_\ell^{(M,i)}$, $\ell = 1, 2, \ldots$; the ℓ^{th} active and idle periods combine to form the ℓ^{th} cycle which has length $\tau_\ell^{(i)} + \theta_\ell^{(M,i)}$. The rvs $\{\tau_\ell^{(i)}, \theta_\ell^{(M,i)}, = 1, 2, \ldots, M; \ell = 1, 2, \ldots\}$ are assumed mutually independent with (i) the rvs $\{\tau_\ell^{(i)}, = 1, 2, \ldots, M; \ell = 1, 2, \ldots\}$ i.i.d. Pareto rvs with parameter α, as in (4.3), with $1 < \alpha < 2$, while (ii) the rvs $\{\theta_\ell^{(M,i)}, = 1, 2, \ldots, M; \ell = 1, 2, \ldots\}$ are finite mean i.i.d. rvs with an arbitrary distribution on $\{1, 2, \ldots\}$ which depends on M. We write

$$a_\tau = \mathbf{E}[\tau] \quad \text{and} \quad a_\theta^{(M)} = \mathbf{E}\left[\theta^{(M)}\right] \tag{6.1}$$

where τ (resp. $\theta^{(M)}$) denotes a generic variable distributed like the i.i.d. rvs $\{\tau_\ell^{(i)}, = 1, 2, \ldots, M; \ell = 1, 2, \ldots\}$ (resp. $\{\theta_\ell^{(M,i)}, = 1, 2, \ldots, M; \ell = 1, 2, \ldots\}$).

Under these assumptions, with each source i, $i = 1, \ldots, M$, we can associate a renewal process $\omega^{(M,i)}$ which counts the cycles of the source i. Alternatively, the process $\omega^{(M,i)}$ is the renewal process where the beginning of cycles for source i are the renewal points; if $a_\tau + a_\theta^{(M)} < \infty$, then the renewal process $\omega^{(M,i)}$ can be constructed to be stationary. Let $\omega^{(M)}$ denote the process obtained by aggregating the M independent stationary renewal processes $\omega^{(M,i)}$, $i = 1, \ldots, M$, and consider the number $\xi_t^{(M)}$ of periods which become active among these M sources at the beginning of slot $[t, t+1)$, $t = 0, 1, \ldots$. The aggregate traffic intensity $\lambda(M)$ is given by

$$\lambda(M) \equiv \frac{M}{a_\tau + a_\theta^{(M)}}. \tag{6.2}$$

We can think of the aggregate process $\omega^{(M)}$ in the following way: We first label active periods in non–decreasing order according to the epoch at which they begin; active periods that begin in the same time slot are ordered according to their source number. With this labeling, for $s = 1, , \ldots,$ we denote by $\omega_s(M)$ the beginning of the s^{th} active period and by $\tau_s(M)$ its length.

In [15], the following asymptotic regime was discussed: Assume that the rvs

$\{\theta^{(M)},\ M = 1, 2, \ldots\}$ are chosen so that

$$\lim_{M \to \infty} \mathbf{P}\left[\theta^{(M)} \leq t\right] = 0, \quad t = 1, 2, \ldots \tag{6.3}$$

while the aggregate traffic intensity remains unchanged, i.e., $\lambda^{(M)} = \lambda$ for some given value $\lambda > 0$, Then, the convergence

$$\{\xi_t^{(M)},\ t = 0, 1, \ldots\} \Longrightarrow_M \{\xi_t,\ t = 0, 1, \ldots\} \tag{6.4}$$

takes place, where the rvs $\{\xi_t,\ t = 0, 1, \ldots\}$ form a sequence of i.i.d. rvs which are distributed according to a Poisson rv with parameter λ. Moreover,

$$\{(\omega_s^{(M)}, \tau_s(M)),\ s = 1, 2, \ldots\} \Longrightarrow_M \{(\omega_s, \tau_s),\ s = 1, 2, \ldots\} \tag{6.5}$$

where the rvs $\{\omega_s,\ s = 1, 2, \ldots\}$ are independent of the i.i.d. rvs $\{\tau_s,\ s = 1, 2, \ldots\}$ which are Pareto distributed according to (4.3). Obviously, in the limiting system we have

$$\xi_t \equiv \sum_{s=1}^{\infty} \mathbf{1}\left[\omega_s = t\right], \quad t = 0, 1, \ldots \tag{6.6}$$

and mapping an active source into a customer, we can interpret ω_s as the arrival epoch of the s^{th} customer and the duration τ_s of its active period as its service time, with service being a metaphor for cell generation. Because a source keeps on generating cells during its activity period, it can be put in one–to–one correspondence with a server dedicated to it. Equipped with this translation, the reader will readily see that the bivariate sequence $\{(\omega_s, \tau_s),\ s = 1, 2, \ldots\}$ does indeed describe the $M|G|\infty$ input model of Section 4. This identification of the $M|G|\infty$ model of Cox as the limiting regime of a large number of on–off sources might help explain its success in modeling packet traffic stream in certain applications [20].

We conclude this section with the following clever observation made by Likhanov, Tsybakov and Georganas in [15]: If attention shifts from the number of cells in buffer to the number of *active sources* in buffer, then the corresponding queuing behavior is described in effect by a discrete–time $M|G|1$ queue. New customers (i.e., new sources) arrive according to the discrete–time "Poisson" process $\{\xi_t,\ t = 0, 1, \ldots\}$ and each customer requires a service of duration which is Pareto distributed according to (4.3); the service discipline is FIFO over the sources but not necessarily on the cells generated. In particular, let $\{N_k,\ k = 1, 2, \ldots\}$ denote the process describing the number of active sources which remain in the buffer at the completion of successive source transmissions. Its steady–state behavior can be obtained by the usual z–transform methods, and yields asymptotics in the form of a power law. However, it should be stressed that this result, while undoubtedly related to that obtained in the previous section, does not capture in a straightforward manner the queuing behavior at the multiplexer in terms of number of cells in buffer, namely (2.1) under the $M|G|\infty$ input.

7 Asymptotics for Fractional Gaussian Noise input

It is of interest to compare these results with that which would be obtained in the situation when the driving sequence is produced by Fractional Gaussian Noise. Recall that the $\mathbb{R}$–valued process $\{n_t,\ t=1,2,\ldots\}$ is a zero–mean FGN process with parameter H, $0<H<1$, if it is a zero–mean Gaussian process with stationary increments and covariance structure

$$\mathrm{cov}\,[n_t,n_s] \equiv \frac{\sigma^2}{2}\left(t^{2H}+s^{2H}-\mid t-s\mid^{2H}\right), \quad s,t=1,2,\ldots \tag{7.1}$$

We consider the recursion with driving sequence $\{b_{t+1},\ t=0,1,\ldots\}$ given by

$$b_{t+1}\equiv n_{t+1}-n_t+r_{in}. \quad t=0,1,\ldots \tag{7.2}$$

with the convention $n_0=0$. This traffic model can be interpreted as the discrete–time analog of that introduced in continuous–time by Norros in [17]. For all $t=1,2,\ldots$, we note that the rv S_t-ct coincides with $n_t+(r_{in}-c)t$, and is therefore a Gaussian rv with mean $(r_{in}-c)t$ and variance σ^2t^{2H}. Hence,

$$\begin{aligned}\Lambda_t(\theta) &= \frac{1}{v_t}\ln\mathbf{E}\left[\exp(\theta v_t\frac{S_t-ct}{a_t})\right]\\ &= \frac{1}{v_t}\ln\mathbf{E}\left[\exp(\theta v_t\frac{n_t+(r_{in}-c)t}{a_t})\right]\\ &= \frac{1}{2}\frac{\sigma^2\theta^2}{v_t}\left(\frac{v_t}{a_t}\right)^2 t^{2H}+\frac{1}{v_t}\frac{v_t}{a_t}\theta(r_{in}-c)t\\ &= \frac{1}{2}\sigma^2\theta^2v_t\left(\frac{t^H}{a_t}\right)^2+\frac{t}{a_t}\theta(r_{in}-c), \quad \theta\in\mathbb{R}\end{aligned} \tag{7.3}$$

and a direct inspection suggests taking $a_t=t$ and $v_t=t^{2(1-H)}$ for all $t=1,2,\ldots$. It is easy to check that $v_{a^{-1}(t/c)}=[t/c]^{2(1-H)}$, thereby dictating the choice $h(b)=b^{2(1-H)}$ and $g(y)=y^{-2(1-H)}$ for all $y>0$. Therefore,

$$\Lambda(\theta)=\Lambda_t(\theta)=\frac{\sigma^2\theta^2}{2}+\theta(r_{in}-c), \quad \theta\in\mathbb{R} \tag{7.4}$$

for all $t = 1, 2, \ldots$, and after some straightforward calculations, we find that

$$\begin{aligned}
\Lambda^\star(z) &= \sup_{\theta \in \mathbb{R}} \{\theta z - \Lambda(\theta)\} \\
&= \sup_{\theta \leq \alpha - 1} \left\{\theta z - (\frac{1}{2}\sigma^2\theta^2 + \theta(r_{in} - c))\right\} \\
&= \sup_{\theta \leq \alpha - 1} \left\{\theta(z - (r_{in} - c)) - \frac{1}{2}\sigma^2\theta^2\right\} \\
&= \frac{1}{2\sigma^2}(z + (c - r_{in}))^2, \quad z \in \mathbb{R}. \qquad (7.5)
\end{aligned}$$

By the Gärtner–Ellis Theorem [6] the process $\{t^{-1}(S_t - ct),\ t = 1, 2, \ldots\}$ satisfies the Large Deviations Principle with good rate function $\Lambda^\star$ under scaling $t^{2(1-H)}$, and the assumptions of Proposition 3.2 both hold.

The rate function $\Lambda^\star$ is continuous on $\mathbb{R}$, and from the remarks around (3.14) we conclude that

$$\begin{aligned}
\gamma^\star = \inf_{z>0} g(z)\Lambda^\star(z) &= \frac{1}{2\sigma^2} \inf_{z>0} z^{-2(1-H)}(z + (c - r_{in}))^2 \\
&= \frac{1}{2\sigma^2} \inf_{z>0} \left(\frac{z + (c - r_{in})}{z^{(1-H)}}\right)^2 \\
&= \frac{1}{2\sigma^2} \left(\frac{(c - r_{in})^H}{C_H}\right)^2 \qquad (7.6)
\end{aligned}$$

where the constant C_H is given by

$$C_H \equiv H^H(1 - H)^{1-H}. \qquad (7.7)$$

Hence, (3.11) now becomes

$$\liminf_{b \to \infty} \frac{1}{b^{2(1-H)}} \ln \mathbf{P}\left[q_\infty > b\right] \geq -\frac{1}{2\sigma^2} \left(\frac{(c - r_{in})^H}{C_H}\right)^2. \qquad (7.8)$$

In fact, for the FGN input model considered here it can be shown that the lower bound is tight, so that

$$\lim_{b \to \infty} \frac{1}{b^{2(1-H)}} \ln \mathbf{P}\left[q_\infty > b\right] = -\frac{1}{2\sigma^2} \left(\frac{(c - r_{in})^H}{C_H}\right)^2. \qquad (7.9)$$

This can be done by applying the results of [7], or alternatively, by a direct argument that relies on the Gaussian nature of the rvs involved. The limit (7.9) can be written in equivalent form as

$$\mathbf{P}\left[q_\infty > b\right] = \exp\left(-\frac{1}{2\sigma^2} \left(\frac{(c - r_{in})^H}{C_H}\right)^2 b^{2(1-H)}(1 + o(1))\right) \quad (b \to \infty) \qquad (7.10)$$

and the decay is therefore Weibullian as announced in [3].

8 Outline of the proof of Proposition 5.1

.

For the $M|GI|\infty$ input model, with $a_t = t$ and $v_t = \ln t$, we want to evaluate $\Lambda_b(\theta)$ as given by (5.3) for all θ in $\mathbb{R}$. The point of departure of our calculations is the key decomposition

$$b_t = b_t^{(0)} + b_t^{(a)}, \quad t = 1, 2, \ldots \tag{8.1}$$

where the rvs $b_t^{(0)}$ and $b_t^{(a)}$ describe the contributions to the number of customers in the system at the beginning of slot [t,t+1) from those initially present (at t = 0) and from the new arrivals, respectively. In fact, with the notation of Section 4 the rv $b_t^{(0)}$ can be expressed as

$$b_t^{(0)} = \sum_{i=1}^{b} \mathbf{1}\left[\sigma_{0,i} > t\right] \tag{8.2}$$

where the i.i.d. rvs $\{\sigma_{0,i},\ i = 1, 2, \ldots\}$ are independent of the rv b which is Poisson distributed with parameter $\lambda\mathbf{E}\left[\sigma\right]$. The rvs $b_t^{(0)}$ and $b_t^{(a)}$ being independent, we find that

$$\Lambda_t^{(b)}(\theta) = \frac{1}{v_t}\Lambda_t^{(0)}(\theta) + \frac{1}{v_t}\Lambda_t^{(a)}(\theta), \quad \theta \in \mathbb{R} \tag{8.3}$$

for all $t = 1, 2, \ldots$ where we have set

$$\Lambda_t^{(0)}(\theta) = \ln \mathbf{E}\left[\exp(\theta \sum_{s=1}^{t} b_s^{(0)})\right], \quad \theta \in \mathbb{R} \tag{8.4}$$

and

$$\Lambda_t^{(a)}(\theta) = \ln \mathbf{E}\left[\exp(\theta \sum_{s=1}^{t} b_s^{(a)})\right], \quad \theta \in \mathbb{R} \tag{8.5}$$

In [18], [19], tedious calculations lead the authors to the following expressions.

Lemma 8.1 *For each $t = 1, 2, \ldots$, we have the expressions*

$$\Lambda_t^{(0)}(\theta) = -\lambda\mathbf{E}\left[\sigma\right]\left(1 - \mathbf{E}\left[\exp(\theta \min(t, \sigma - 1))\right]\right) \tag{8.6}$$

and

$$\begin{aligned} \Lambda_t^{(a)}(\theta) &= -\lambda\mathbf{E}\left[(t-\sigma)^+(1 - \exp\theta\sigma) + \min(t,\sigma)\right] \\ &\quad + \lambda\left(1 - e^{-\theta}\right)^{-1}\mathbf{E}\left[\exp(\theta \min(t,\sigma)) - 1\right] \end{aligned} \tag{8.7}$$

for all θ in $\mathbb{R}$.

The proof of Lemma 8.1 is omitted in the interest of brevity. The conclusions of Lemma 8.1 are valid regardless of the pmf G assumed for σ. To proceed further, we specify G to be as described in (4.4). Using this choice in (8.6)–(8.7) leads to great simplifications in the expressions as we shall now see: For each $\beta > 0$, we define

$$F_\beta(t,\theta) = \sum_{r=1}^{t} r^{-\beta} e^{\theta r}, \quad \theta \in \mathbb{R} \tag{8.8}$$

for all $t = 1, 2, \ldots$; the identity

$$\sum_{r=1}^{t} \left(r^{-\beta} - (r+1)^{-\beta}\right) e^{\theta r} = F_\beta(t,\theta)(1-e^{-\theta}) + 1 - e^{\theta t}(t+1)^{-\beta}, \quad \theta \in \mathbb{R} \tag{8.9}$$

can be easily confirmed by simple algebra. Making use of this elementary fact we readily obtain the following relations from Lemma 8.1:

Lemma 8.2 *For each $t = 1, 2, \ldots$, the expression*

$$\Lambda_t^{(0)}(\theta) = \lambda \mathbf{E}[\sigma](1-e^{-\theta})\left(e^{-\theta} F_\alpha(t,\theta) + e^{\theta t}\mathbf{P}[\sigma > t] - 1\right) \tag{8.10}$$

holds for all θ in $\mathbb{R}$.

Proof. Substituting (4.4) into (8.6), we see that

$$\begin{aligned} \Lambda_t^{(0)}(\theta) &= -\lambda \mathbf{E}[\sigma]\left(1 - \mathbf{E}[\exp(\theta \min(t, \sigma - 1))]\right) \\ &= -\lambda \mathbf{E}[\sigma]\left(1 - e^{-\theta}\sum_{r=1}^{t} g_r e^{\theta r} + e^{\theta t}\mathbf{P}[\sigma > t]\right) \\ &= \lambda \mathbf{E}[\sigma](1-e^{-\theta})\left(e^{-\theta}F_\alpha(t,\theta) + e^{\theta t}\mathbf{P}[\sigma > t] - 1\right) \end{aligned}$$

where in the last step we have used (8.9) with $\beta = \alpha$. ■

Lemma 8.3 *For each $t = 1, 2, \ldots$, the expression*

$$\Lambda_t^{(a)}(\theta) = \lambda(1-e^{-\theta})\left((t+1)F_\alpha(t,\theta) - F_{\alpha-1}(t,\theta)\right) \tag{8.11}$$

holds for all θ in $\mathbb{R}$.

Proof. This time, substitution of (4.4) into (8.7) yields

$$\begin{aligned}\Lambda_t^{(a)}(\theta) &= -\lambda\mathbf{E}\left[(t-\sigma)^+\left(1-\exp(\theta\sigma)\right)+\min\left(t,\sigma\right)\right] \\ &\quad +\lambda\left(1-e^{-\theta}\right)^{-1}\mathbf{E}\left[\exp(\theta\min(t,\sigma))-1\right] \\ &= -\lambda t+\lambda\sum_{r=1}^{t}(t-r)g_r e^{\theta r} \\ &\quad +\lambda\left(1-e^{-\theta}\right)\left(\sum_{r=1}^{t}g_r e^{\theta r}+\mathbf{P}\left[\sigma>t\right]e^{\theta t}-1\right).\end{aligned}$$

Using (8.9) into this last expression and simplifying, we readily obtain the desired result. ■

It is now time to substitute in (8.10)–(8.11) for θ by $\theta_t \equiv \frac{\theta v_t}{t}$, and to take note of the fact that

$$\lim_{t\to\infty}\frac{1-e^{\theta_t}}{\theta_t}=1. \tag{8.12}$$

This elementary fact already allows us to conclude to the asymptotic equivalences

$$\frac{\Lambda_t^{(0)}(\theta_t)}{v_t}\sim L_t^{(0)}(\theta_t) \quad\text{and}\quad \frac{\Lambda_t^{(a)}(\theta_t)}{v_t}\sim L_t^{(a)}(\theta_t), \quad \theta\in\mathbb{R} \tag{8.13}$$

where for all $t=1,2,\ldots$ and θ in $\mathbb{R}$, we have defined

$$L_t^{(0)}(\theta_t)\equiv\frac{\lambda\theta}{t}\mathbf{E}\left[\sigma\right]\left(e^{\theta_t t}\mathbf{P}\left[\sigma>t\right]+e^{-\theta_t}F_\alpha(t,\theta_t)-1\right) \tag{8.14}$$

and

$$L_t^{(a)}(\theta_t)\equiv\frac{\lambda\theta}{t}\left((t+1)F_\alpha(t,\theta_t)-F_{\alpha-1}(t,\theta_t)\right). \tag{8.15}$$

In short, combining (8.3) and (8.13), we have

$$\Lambda_t^b(\theta_t)\sim L_t^{(0)}(\theta_t)+L_t^{(a)}(\theta_t), \quad \theta\in\mathbb{R}. \tag{8.16}$$

The next step consists in getting some insight into the asymptotic behavior of $L_t^{(0)}(\theta_t)$ and $L_t^{(a)}(\theta_t)$, through that of $F_\alpha(t,\theta_t)$ and $F_{\alpha-1}(t,\theta_t)$, as $t\to\infty$. This is the content of the next two lemmas; their proofs are given in the Appendix.

Lemma 8.4 *For each $\beta>0$, the asymptotics*

$$\lim_{t\to\infty}\frac{F_\beta(t,\theta_t)}{t}=\begin{cases}0 & \textit{if}\ \theta\leq\beta\\ \infty & \textit{if}\ \theta>\beta\end{cases} \tag{8.17}$$

hold.

Lemma 8.5 *The convergence*

$$\lim_{t\to\infty} \frac{1}{t}\left((t+1)F_\alpha(t,\theta_t) - F_{\alpha-1}(t,\theta_t)\right) = \begin{cases} \mathbf{E}\,[\sigma] & \textit{if}\ \theta \le \alpha - 1 \\ \infty & \textit{if}\ \theta > \alpha - 1 \end{cases} \tag{8.18}$$

holds.

Lemma 8.4 can be rephrased as saying that

$$\lim_{t\to\infty} L_t^{(0)}(\theta_t) = \begin{cases} 0 & \text{if } \theta \le \alpha \\ \infty & \text{if } \theta > \alpha \end{cases} \tag{8.19}$$

whereas Lemmas 8.4 and 8.5 together imply

$$\lim_{t\to\infty} L_t^{(a)}(\theta_t) = \begin{cases} \lambda\theta\mathbf{E}\,[\sigma] & \text{if } \theta \le \alpha - 1 \\ \infty & \text{if } \theta > \alpha - 1 \end{cases} \tag{8.20}$$

and the desired conclusion (5.3) follows via (8.16).

9 Appendix – Proofs

In both proofs, Γ denotes an element of (0,1) to be selected suitably during the discussion.

Proof of Lemma 8.4. Two cases naturally emerge, namely $\theta \le \beta$ and $\theta > \beta$.

a. Assume $\theta \le \beta$: Write $\beta_t = \beta\frac{v_t}{t}$ for all $t = 1, 2, \ldots$. In view of the inequalities

$$0 < F_\beta(t,\theta_t) \le F_\beta(t,\beta_t), \quad \theta \le \beta,\ t = 1,2,\ldots \tag{9.1}$$

the first part of Lemma 8.4 will be established, provided we show

$$\limsup_{t\to\infty} \frac{F_\beta(t,\beta_t)}{t} \le 0. \tag{9.2}$$

With this in mind we compute an upper bound for $F_\beta(t,\beta_t)$, as follows:

$$\begin{aligned} F_\beta(t,\beta_t) &= \sum_{r=1}^{[\Gamma t]} r^{-\beta} e^{\frac{\beta v_t r}{t}} + \sum_{r=[\Gamma t]+1}^{t} r^{-\beta} e^{\frac{\beta v_t r}{t}} \\ &\le e^{\frac{\beta v_t [\Gamma t]}{t}} \sum_{r=1}^{[\Gamma t]} r^{-\beta} + ([\Gamma t]+1)^{-\beta} \sum_{r=[\Gamma t]+1}^{t} e^{\frac{\beta v_t r}{t}} \\ &\le e^{\frac{\beta v_t [\Gamma t]}{t}} \left(1 + \int_1^{[\Gamma t]} x^{-\beta} dx\right) \\ &\quad + \ ([\Gamma t]+1)^{-\beta} e^{\frac{\beta v_t ([\Gamma t]+1)}{t}} \sum_{r=0}^{t-([\Gamma t]+1)} e^{\frac{\beta v_t r}{t}}. \end{aligned}$$

Simplifying, dividing by t and letting t go to infinity readily yield

$$\limsup_{t\to\infty} \frac{F_\beta(t,\beta_t)}{t} \leq \limsup_{t\to\infty} \frac{\beta}{\beta-1} t^{\Gamma\beta-1} \tag{9.3}$$

and (9.2) is obtained if we select Γ in the interval $(0, \min(\beta^{-1}, 1))$.

b. Assume $\theta > \beta$: Again, splitting the expression of interest as before, we see by easy bounding arguments that

$$\begin{aligned} F_\beta(t,\theta_t) &= \sum_{r=1}^{[\Gamma t]} r^{-\beta} e^{\frac{\theta v_t r}{t}} + \sum_{r=[\Gamma t]+1}^{t} r^{-\beta} e^{\frac{\theta v_t r}{t}} \\ &\geq \sum_{r=[\Gamma t]+1}^{t} r^{-\beta} e^{\frac{\theta v_t r}{t}} \\ &\geq e^{\frac{\theta v_t [\Gamma t]}{t}} \int_{[\Gamma t]+1}^{t+1} x^{-\beta} dx \\ &= t^{\frac{\theta[\Gamma t]}{t}-\beta+1} \frac{(\frac{[\Gamma t]+1}{t})^{-\beta+1} - (1+\frac{1}{t})^{-\beta+1}}{\beta-1}. \end{aligned} \tag{9.4}$$

Dividing by t on both sides of (9.4) and then taking lim inf give

$$\liminf_{t\to\infty} \frac{1}{t} F_\beta(t,\theta_t) \geq \liminf_{t\to\infty} t^{\theta\Gamma-\beta} \frac{\Gamma^{-\beta+1}-1}{\beta-1}. \tag{9.5}$$

Under the constraint $\theta > \beta$ we can always select Γ in the interval $(\frac{\beta}{\theta}, 1)$, thereby leading to $\liminf_{t\to\infty} \frac{1}{t} F_\beta(t,\theta_t) = \infty$ ■

Proof of Lemma 8.5. As before two cases need to be considered separately:

a. Assume $\theta \leq \alpha - 1$: Lemma 8.4 (with $\beta = \alpha - 1$), already implies the asymptotic equivalence

$$\frac{1}{t}\left((t+1)F_\alpha(t,\theta_t) - F_{\alpha-1}(t,\theta_t)\right) \sim F_\alpha(t,\theta_t) \tag{9.6}$$

Next, we observe that

$$F_\alpha(t,\theta_t) - \mathbf{E}[\sigma] = \sum_{r=1}^{t} r^{-\alpha}(e^{\frac{\theta v_t r}{t}} - 1) \geq 0, \quad t = 1, 2, \ldots \tag{9.7}$$

and we shall show shortly that

$$\limsup_{t\to\infty} F_\alpha(t,\theta_t) - \mathbf{E}[\sigma] \leq 0. \tag{9.8}$$

It will then be plain that $\lim_{t\to\infty} F_\alpha(t,\theta_t) = \mathbf{E}[\sigma]$, and combining this fact with (9.6) we readily get the first part of Lemma 8.5.

To establish (9.8) we begin by noting that

$$\begin{aligned}
F_\alpha(t,\theta_t) - \mathbf{E}[\sigma] &\le \frac{\theta v_t}{t}\sum_{r=1}^{[\Gamma t]} r^{-\alpha+1} e^{\frac{\theta v_t r}{t}} + \sum_{r=[\Gamma t]+1}^{t} r^{-\alpha}(e^{\frac{\theta v_t r}{t}} - 1) \\
&\le \frac{\theta v_t}{t} e^{\frac{\theta v_t [\Gamma t]}{t}} \left(1 + \int_1^{[\Gamma t]} x^{-\alpha+1} dx\right) \\
&\quad + \ ([\Gamma t]+1)^{-\alpha} \int_{[\Gamma t]+1}^{t+1} (e^{\frac{\theta v_t x}{t}} - 1) dx.
\end{aligned}$$

Simplifying and taking the limsup on both sides of the inequality readily give (9.8).

b. Assume $\alpha - 1 < \theta$: We note that

$$\begin{aligned}
(t+1)F_\alpha(t,\theta_t) - F_{\alpha-1}(t,\theta_t) &= \sum_{r=1}^{t} r^{-\alpha} e^{\frac{\theta v_t r}{t}} (t+1) - \sum_{r=1}^{t} r^{-\alpha+1} e^{\frac{\theta v_t r}{t}} \\
&= \sum_{r=1}^{[\Gamma t]} r^{-\alpha} e^{\frac{\theta v_t r}{t}} (t+1-r) \\
&\quad + \sum_{r=1+[\Gamma t]}^{t} r^{-\alpha} e^{\frac{\theta v_t r}{t}} (t+1-r)) \\
&\ge \sum_{r=1}^{[\Gamma t]} r^{-\alpha} e^{\frac{\theta v_t r}{t}} (t+1-r) \\
&\ge (t+1-[\Gamma t])([\Gamma t])^{-\alpha} \sum_{r=1}^{[\Gamma t]} e^{\frac{\theta v_t r}{t}}.
\end{aligned}$$

Letting t go to infinity in the above, we conclude that

$$\begin{aligned}
&\liminf_{t\to\infty} \frac{1}{t}((t+1)F_\alpha(t,\theta_t) - F_{\alpha-1}(t,\theta_t)) \\
\ge &\liminf_{t\to\infty} \frac{\Gamma^{-\alpha} t^{1-\alpha}}{\theta v_t}(1 - \Gamma + \frac{1}{t})(t^{\Gamma\theta} - 1) = \infty
\end{aligned}$$

for any value of Γ in the interval $(\frac{\alpha-1}{\theta}, 1)$. ∎

References

[1] R.G. Addie, M. Zukerman and T. Neame, "Fractal traffic: Measurements, modeling and performance evaluation in *Proceedings of Infocom '95*, Boston (MA), April 1995, pp. 985–992.

[2] J. Beran, *Statistics for Long-Memory Processes*, Chapman and Hall, New York (NY), 1994.

[3] J. Beran, R. Sherman, M. S. Taqqu and W. Willinger "Long-range dependence in variable bit–rate video traffic," *IEEE Transactions on Communications* **COM–43** (1995), pp. 1566–1579.

[4] D. R. Cox, "Long–Range Dependence: A Review," *Statistics: An Appraisal*, H. A. David and H. T. David, Eds., The Iowa State University Press, Ames (IA), 1984, pp 55–74.

[5] D. R. Cox and V. Isham, *Point Processes*, Chapman and Hall, New York (NY), 1980.

[6] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications*, Jones and Bartlett, Boston (MA), 1993.

[7] N. G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single server queue, with applications," Technical Report **DIAS-STP-93-30**, Dublin Institute for Advanced Studies, Dublin (Ireland), 1993. *Proceedings of the Cambridge Philosophical Society*, forthcoming 1995.

[8] A. Erramilli, O. Narayan and W. Willinger, "Experimental queuing analysis with long–range dependent packet traffic," Preprint 1994.

[9] H. J. Fowler and W. E. Leland, "Local area network traffic characteristics, with implications for broadband network congestion management," *IEEE Journal on Selected Areas in Communications* **JSAC–9**(1991), pp. 1139–1149.

[10] M. Garrett and W. Willinger, "Analysis, modeling and generation of self–similar VBR video traffic," *Proceedings of SIGCOMM '94*, September 1994, pp. 269–280.

[11] P.W. Glynn and W. Whitt, "Logarithmic asymptotics for steady–state tail probabilities in a single–server queue," *Journal of Applied Probability* (1993).

[12] R. Guérin, H. Ahmadi and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high–speed networks," *IEEE Journal on Selected Areas in Communications* **JSAC–9** (1991), pp. 968–981.

[13] G. Kesidis, J. Walrand and C.S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Transactions on Networking* **1** (1993), pp. 424–428.

[14] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self–similar nature of ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking* **2** (1994), pp. 1–15.

[15] N. Likhanov, B. Tsybakov and N.D. Georganas, "Analysis of an ATM buffer with self–similar ("fractal") input traffic," in *Proceedings of Infocom '95*, Boston (MA), April 1995, pp. 985–992.

[16] R.M. Loynes, "The stability of a queue with non–independent inter–arrival and service times," *Proceedings of the Cambridge Philosophical Society* **58** (1962), pp. 497–520.

[17] I. Norros, "A storage model with self–similar input," *Queuing Systems – Theory & Applications*,**16** (1994), pp. 387-396.

[18] M. Parulekar, *Buffer Engineering for Self–Similar Traffic*, Ph.D. Thesis, Electrical Engineering Department, University of Maryland, College Park (MD). Expected December 1996.

[19] M. Parulekar and A.M. Makowski, "Buffer overflow probabilities for $M|G|\infty$ input processes," in preparation.

[20] V. Paxson and S. Floyd, "Wide area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking* **3** (1993), pp. 226–244.

[21] W. Whitt, "Tail probability with statistical multiplexing and effective bandwidths in multi–class queues," *Telecommunication Systems*, to appear.

[22] W. Willinger, M. S. Taqqu, W. E. Leland and D. V. Wilson, "Self–similarity in high–speed packet traffic: Analysis and modeling of ethernet traffic measurements," *Statistical Science*, to appear.