Impact of Bursty Traffic on Queues

Philippe Nain

INRIA
B.P. 93, 06902 Sophia Antipolis Cedex, France
nain@sophia.inria.fr

Published in:
Statistical Inference for Stochastic Processes, Vol. 5, pp. 307-320, 2002.

Abstract

The impact of bursty traffic on queues is investigated in this paper. We consider a discrete-time single server queue with an infinite storage room, that releases customers at the constant rate of c customers/slot. The queue is fed by an M/G/∞ process. The M/G/∞ process can be seen as a process resulting from the superposition of infinitely many "sessions": sessions become active according to a Poisson process; a station stays active for a random time,with probability distribution G, after which it becomes inactive. The number of customers entering the queue in the time-interval $[t, t+1)$ is then defined as the number of active sessions at time t ($t = 0, 1, \ldots$) or, equivalently, as the number of busy servers at time t in an M/G/∞ queue, thereby explaining the terminology. The M/G/∞ process enjoys several attractive features: First, it can display various forms of dependencies, the extent of which being governed by the service time distribution G. The heavier the tail of G, the more bursty the M/G/∞ process. Second, this process arises naturally in teletraffic as the limiting case for the aggregation of on/off sources [27]. Third, it has been shown to be a good model for various types of network traffic, including telnet/ftp connections [37] and variable-bit-rate (VBR) video traffic [24]. Last but not least, it is amenable to queueing analysis due to its very strong structural properties. In this paper we compute an asymptotic lower bound for the tail distribution of the queue length. This bound suggests that the queueing delays will dramatically increase as the burstiness of the M/G/∞ input process increases. More specifically, if the tail of G is heavy, implying a bursty input process, then the tail of the queue length will also be heavy. This result is in sharp contrast with the exponential decay rate of the tail distribution of the queue length in presence of "non-bursty" traffic (e.g. Poisson-like traffic).

Keywords: Self-similar process; long-range dependence; subexponential distribution; queue; performance evaluation; data network.

1 Introduction

Recent measurements [11, 25] have shown that data traffic in networks (e.g. Internet) may exhibit "similar looking behavior" over an an extremely wide range of time scales (from a few milliseconds to several hours). These observations are in sharp contrast to the Poisson-like nature of traditional telephone and data traffic (like in the Arpanet, the "ancestor" of the Internet – [23]). As a result of these findings, one now speaks of the bursty/self-similar/fractal nature of the traffic and discusses the failure of Poisson modeling [37]. At this time, both the reasons why self-similar-like traffic – referred to as bursty traffic from now on – builds up in the network and the impact of these traffic on the performance of the network (delay, loss probability, throughput, etc.) are still the object of ongoing research as well as debates!

While electronic and software breakdowns may still occur in today's complex networks, in the vast majority of cases the degradation in the performance results from the congestion in the network. When congestion occurs, queues in routers build up to create long delays and losses, that will in turn decrease the quality of service (QoS) provided to the users. In the case of "conventional" traffic (eg. Poisson-like traffic) packet losses due to congestion are expected to have less impact on the QoS than in the case of bursty traffic where a burst of packets arriving at a full router is likely to be partially or entirely lost, a situation that, when repeated, may harm the performance.

The objective of this paper is to quantify the impact of bursty traffic on the performance in the framework of queueing theory. To this end, we will consider a discrete-time single-server queue with a constant release rate of c customers/slot and fed by an "M/G/∞ process". Under the assumption that the queue is stable, we will evaluate the tail distribution of the stationary queue length q. We will actually content ourselves with the computation of a lower bound on $\mathbf{P}(q > x)$ when $x \to \infty$. This bound will already tell us a lot on the impact of burstiness on performance.

The M/G/∞ process is a process $\{b_t, t = 0, 1, \ldots\}$ that counts the number of busy servers at times $t = 0, 1, \ldots$ in an M/G/∞ queue [42]. A precise definition of the M/G/∞ process is given in Section 4. It was first mentioned by Cox and Isham [10] as an instance of a process exhibiting long-range dependence, which occurs when the service time distribution G in the M/G/∞ queue is Pareto (with parameter lying in $(1, 2)$ – see Section 4). The M/G/∞ process enjoys several attractive features. First, it can display various forms of dependencies, the extent of which is governed by the service time distribution G (see (4.2)). Second, this process arises naturally in teletraffic as the limiting case for the aggregation of on/off sources [27]. Third, it has been shown to be a good model for various types of network traffic, including telnet and ftp connections [37] and variable-bit-rate (VBR) video traffic [24]. Finally, it is amenable to various queueing analysis thanks to its very strong structural properties (see Section 4). Other proposed traffic models include fractional Brownian motion and its discrete-time analog, fractional Gaussian noise [1, 15, 31, 32], on-off sources with heavy-tailed activity periods [2, 5, 6, 7, 9, 17, 20, 40] and an aggregation of independent memoryless on/off sources [19]. Already, all these studies have exposed clearly the limitations of traditional traffic models in predicting storage requirements.

The use of the M/G/∞ process as a traffic model was first advocated in [37] where the authors have found that it matches reasonably well some wide area applications. Queueing performance for queues fed by a M/G/∞ process have been reported in [13, 20, 29, 34, 35, 36, 38, 43]. Queueing metrics investigated in these works include the distribution of the queue length as well as the probability of overflow and the distribution of the time to overflow in case of finite buffers.

Through the study of a single server queue we provide in this paper a comprehensive discussion on the impact of bursty traffic on queueing performance. The paper is organized as follows: the main concepts associated with bursty traffic are collected in Section 2; Section 3 reports basic results on the behavior of the queue length in case of non-bursty inputs. M/G/∞ input processes are then introduced and discussed at length in Section 4. We will observe that not only the buffer content is heavy-tailed when the M/G/∞ process is long-range dependent (LRD), but also that its tail may remain (moderately) heavy when it is short-range dependent (SRD). These results are in sharp contrast to the exponential decay rate encountered in Section 3 when the input is a "weakly correlated" process. These results already indicate that self-similarity and LRD are not the only ingredients in the building up of long queueing delays. The paper ends with some remarks on the relevance of bursty traffic to the dimensioning of network resources (Section 5).

A word on the notation in use: $\mathbb{N} = \{0, 1, 2, \ldots\}$ (resp. $\mathbb{N}^\star = \{1, 2, \ldots\}$) will denote the set of all nonnegative (resp. positive) integer numbers. A renewal sequence of random variables (rv's) is a sequence of independent and identically distributed (iid) rv's. For any rv X with cumulative probability distribution $F(x) = \mathbf{P}(X \leq x)$, $\overline{F}(x) = 1 - F(x)$ denotes the probability distribution of its tail. For any real number x, $\lceil x \rceil$ will denote the smallest integer larger than or equal to x. Last, $I(A)$ will denote the indicator function of the event A.

2 Self-similarity, long-range dependence and subexponentiality

Throughout we will only consider discrete-time stochastic sequence $\{X_t, t \in \mathbb{N}\}$. In the networking setting the rv X_t may represent, for instance, the number of packets entering a router in the time-interval (or time-slot) $[t, t+1)$.

A stationary sequence [41] $\mathbf{X} = \{X_t, t \in \mathbb{N}\}$ is *self-similar* with (Hurst) parameter $H \geq 0$ if its distribution coincides with that of the process $\mathbf{X}_m = \{m^{-H}(X_{tm} + \cdots + X_{(t+1)m}), t \in \mathbb{N}\}$ for every $m \in \mathbb{N}^\star$. In other words, a process is self-similar if it is equivalent (in distribution) to the process resulting from summing up the original process over non-overlapping blocks of identical lengths (say m) divided by m^H. For instance, the increments of a fractional Brownian motion are self-similar with parameter H [3, 30].

If $\mathbf{X}$ is self-similar with Hurst parameter H and possesses a finite second-order moment, then its autocorrelation function $r(k) = \mathrm{cov}(X_t, X_{t+k})/\mathrm{var}(X_t)$ is given by [3, Chapter 2]

$$r(k) = (1/2)\left((k+1)^{2H} - 2k^{2H} - (k-1)^{2H}\right), \quad k \in \mathbb{N}^\star.$$

From this identity we deduce the asymptotics

$$r(k) \sim H(2H-1)k^{-2(1-H)} \quad (k \to \infty). \tag{2.1}$$

A stationary sequence $\mathbf{X}$ is called *asymptotically second-order self-similar* if the autocorrelation function $r_m(k)$ of the "aggregated" process $\mathbf{X}_m$, converges to $r(k)$ as m tends to infinity. An instance of such a process is given in Section 4.

A stationary sequence is *long-range dependent* if its autocorrelation function is not summable in L_1, namely if $\sum_{k\geq 0} |r(k)| = \infty$. We see from (2.1) that a self-similar process with parameter $1/2 < H < 1$ is long-range dependent.

Ties exist between self-similar/long-range dependent processes and the class $\mathcal{S}$ of *subexponential* rv's. Formally, a nonnegative rv X is subexponential if $\mathbf{P}(X + X' > x)/\mathbf{P}(X > x) \sim 2$ $(x \to \infty)$, where X' is an independent copy of X [14]. A key consequence of this definition is that subexponentiality means "slower than exponential tails", in the sense that if $X \in \mathcal{S}$ then $\lim_{x\uparrow\infty} e^{\lambda x}\,\mathbf{P}(X > x) = \infty$ for all $\lambda > 0$. Let us briefly see what is behind the definition. Let $\{X_n, n \in \mathbb{N}^\star\}$ be iid rv's and define M_n as the maximum of $X_1, \ldots, X_n$ and S_n as the sum of $X_1, \ldots, X_n$. It is always the case that $\lim_{x\uparrow\infty} \mathbf{P}(M_n > x)/\mathbf{P}(X > x) = n$ and that $\liminf_{x\uparrow\infty} \mathbf{P}(S_n > x)/\mathbf{P}(X > x) \geq n$ as elementary considerations show. What subexponentiality says is that $\mathbf{P}(S_n > x) \sim \mathbf{P}(M_n > x)$ $(x \to \infty)$ for $n = 2, 3, \ldots$ if $X_n \in \mathcal{S}$ (this result flows from the very definition of $\mathcal{S}$). As nicely put by A. M. Makowski during a seminar given at INRIA in the Spring of 1998, subexponentiality can be seen as "conspiracy versus rogue loner".

Examples of subexponential rv's include Pareto, Log-normal and Weibull rv's. A nonnegative rv X is Pareto if $\mathbf{P}(X > x) = L(x)\,x^{-\alpha}$, $\alpha > 1$, where $L(x)$ is a slowly varying function (i.e. $\lim_{x\uparrow\infty} L(tx)/L(x) = 1$ all $t > 0$), Log-normal if $X \stackrel{d}{=} \exp(\delta\, U + \mu)$ with $U \stackrel{d}{=} N(0,1)$, and Weibull if $\mathbf{P}(X > x) = \exp(-ax^\nu)$ with $a > 0$ and $0 < \nu < 1$.

As an example of the existing connections between self-similar processes and subexponential rv's, it has been shown [45] that the superposition of infinitely many, strictly alternating, independent, identical and adequately normalized on/off sources is a fractional Brownian motion. This convergence takes place only for particular heavy-tailed distributions of the length of on and off periods, typically, Pareto-like distributions.

3 Queue under non-bursty traffic

Consider a discrete-time single server queue with an infinite buffer, receiving $b_t \in \mathbb{N}$ packets in the time-interval (slot) $[t, t+1)$ and sending out at most $c \in \mathbb{N}^\star$ packets in every slot. Then q_t, the number of packets in the system at time t, satisfies the Lindley recursion

$$q_{t+1} = \max\{q_t + b_t - c, 0\}, \quad t \in \mathbb{N}. \tag{3.1}$$

If the input process $\{b_t, t \in \mathbb{N}\}$ is stationary and ergodic [41, Chapter V] with finite mean $\rho = \mathbf{E}[b_t] > 0$ and $q_0 \in \mathbb{N}$ (for instance $q_0 = 0$), then q_t converges in distribution to a proper rv q, that is $\mathbf{P}(q_t \leq x) \to \mathbf{P}(q \leq x)$ as $t \to \infty$ for all $x \in \mathbb{N}$. The rv q is called the stationary queue-length.

An explicit expression for $\mathbf{P}(q \leq x)$ is in general not available, even when $\{b_t, t \in \mathbb{N}\}$ is a renewal sequence. Fortunately, *bounds* can easily be obtained in the latter case. Specializing a result by Kingman [22] to a queue fed by the renewal sequence $\{b_t, t \in \mathbb{N}\}$ (with generic element b and common distribution $B(x) = \mathbf{P}(b \leq x)$), we obtain

$$a\, e^{-\theta^* x} \leq \mathbf{P}(q > x) \leq e^{-\theta x}, \quad x \in \mathbb{N} \tag{3.2}$$

for all $0 \leq \theta \leq \theta^* = \sup\{y > 0 : \mathbf{E}[\exp(y(b-c))] = 1\}$ with a a nonnegative constant given by $a = \inf_{x>0}(1 - B(x))/\int_0^x \exp(\theta^*(u-x))\, B(du)$. The upper bound in (3.2) is non-trivial under the stability condition $\mathbf{E}[b] < c$ in that $\theta^* > 0$ under this condition.

The bounds in (3.2) extend to the case when the sequence $\{b_t, t \in \mathbb{N}\}$ is Markov modulated [12, 28], in which case the b_t's are weakly correlated rv's.

We conclude from (3.2), and the extension of this result to Markov modulated input processes as mentioned above, that the complementary distribution of the queue-length decreases *exponentially* fast to zero as x tends to infinity (at least when $a > 0$, which arises for most distributions of practical interest – see a discussion in [28]), more precisely,

$$\lim_{x \to \infty} \frac{1}{x} \log \mathbf{P}(q > x) = -\theta^* \tag{3.3}$$

when $\{b_t, t \in \mathbb{N}\}$ is a Markov modulated sequence. In particular, (3.3) holds when $\{b_t, t \in \mathbb{N}\}$ is a renewal sequence. The validity of (3.3) actually extends much beyond the Markovian setting [8, Thm 3.9], [18, Thm 1], the key condition being that

$$\Phi(y) = \lim_{t \to \infty} \frac{1}{t} \log \mathbf{E}[\exp(y(b_0 + \cdots + b_{t-1}))]$$

exists and is finite in the vicinity of zero.

None of the input sequences considered so far exhibits any kind of self-similar or long-range dependence properties. They all have in common the existence and finiteness of $\Phi(y)$, at least for y in a neighborhood of zero, which yields an exponential decrease of $\mathbf{P}(q > x)$ as x tends to infinity. In the next section we show that a very different decrease takes place when the queue is fed by a bursty process.

4 Queue under bursty traffic

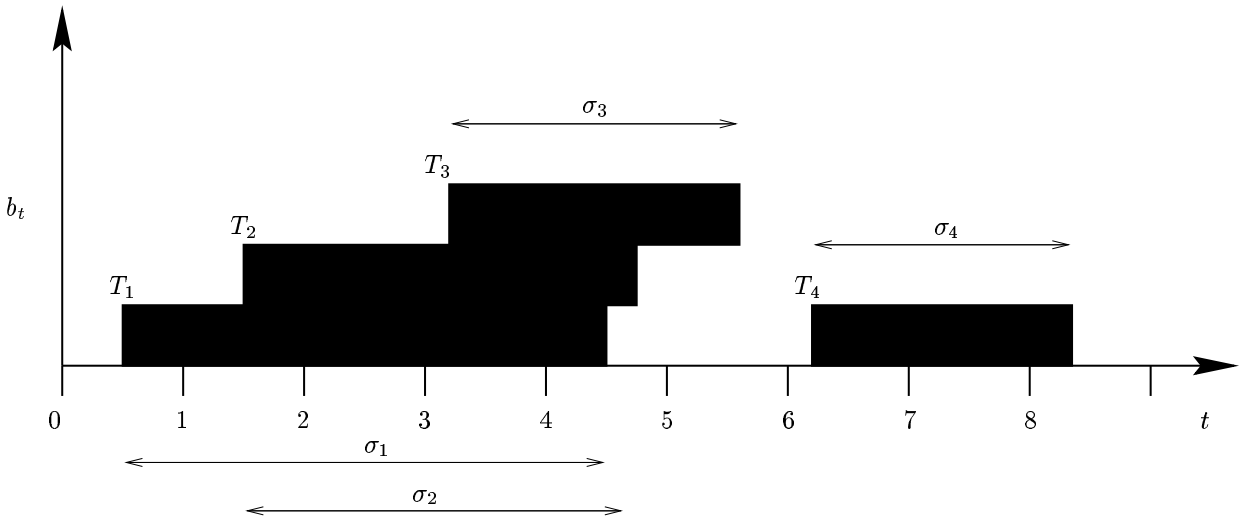
In this section we revisit the discrete-time queueing model introduced in Section 3, this time assuming that the queue is fed by the so-called "M/G/∞ process".

In order to define the M/G/∞ process we introduce a Poisson process $\{T_j, j \in \mathbb{N}^\star\}$ with intensity $\lambda > 0$ and a renewal sequence of rv's $\{\sigma_j, j \in \mathbb{N}^\star\}$, independent of $\{T_j, j \in \mathbb{N}^\star\}$, with generic element σ and common cumulative probability distribution $G(x) = \mathbf{P}(\sigma \leq x)$. Let $\bar{\sigma} = \mathbf{E}[\sigma] < \infty$.

Consider now the discrete-time, integer-valued process $\{b_t, t \in \mathbb{N}\}$ defined as

$$b_t = \sum_{0 \leq T_j < t} I(\sigma_j > t - T_j), \quad t \in \mathbb{N}^\star$$

with $b_0 = 0$ a.s. In the queueing setting b_t can be interpreted as follows: assume that customers arrive at a queue with infinitely many servers at times $0 \leq T_1 < T_2 < \cdots$ and that customer arriving at time T_j ($j \in \mathbb{N}^\star$) carries with it a service time σ_j (this queueing system is known in the literature as the M/G/∞ queue [42]). If we further assume that this M/G/∞ queue is empty when the first customer arrives, then b_t gives the number of busy servers at time t. For that reason, the process $\{b_t, t \in \mathbb{N}\}$ is called an M/G/∞ (input) process.



$b_0 = 0$, $b_1 = 1$, $b_2 = 2$, $b_3 = 2$, $b_4 = 3$, $b_5 = 1$, $b_6 = 0$, $b_7 = 1$, $b_8 = 1$

Figure 1: M/G/∞ process

In a networking setting, the traffic generated by an M/G/∞ process can be seen (cf. Figure 1) as the traffic resulting from the superposition of infinitely many "sessions", each session becoming active at random (Poisson) times and staying active for a random time (with distribution G). During an activity period a session sends one packet per unit time to the queue (router). Therefore, if b_t sessions are simultaneously active at time t, then b_t packets will be sent to the queue in the time-interval $[t, t+1)$. Moreover, if these b_t sessions stay simultaneously active for at least n consecutive time-slots then at least b_t packets will be generated at times $t, t+1, \ldots, t+n$. With this interpretation we see that the heavier the tail of the distribution of the activity periods $\{\sigma_j, j \in \mathbb{N}^\star\}$

the more *bursty* the input process to the queue. In particular, if σ is an heavy-tailed rv (e.g. Pareto rv – see the discussion after the proof of Proposition 4.1) then the traffic will be very bursty.

For later use, we introduce $G_1(x) = (1/\overline{\sigma}) \int_0^x \overline{G}(u)du$ $(x \geq 0)$, the integrated tail distribution of G. In words, G_1 is the distribution of the residual lifetime of a rv observed at an arbitrary instant, a key quantity in renewal theory.

Lemma 4.1 below, whose proof follows from [4, Chapter 6] and [42, pp. 160-162] reports some basic features of the M/G/∞ process that we will use in the proof of the main result in Section 4.

Lemma 4.1 *If $\{b_t, t \in \mathbb{N}\}$ is an M/G/∞ process then the distribution of the sequence $\{b_{t+k}, t \in \mathbb{N}\}$ converges monotonically for $k \to \infty$ to the distribution of a proper stationary and ergodic sequence $\{b^t, t \in \mathbb{N}\}$, with*

$$b^t = \sum_{j=1}^{b^0} I\left(\hat{\sigma}_j > t\right) + \sum_{s=0}^{t-1} \sum_{s \leq T_j < s+1} I(\sigma_j > t - T_j). \tag{4.1}$$

In (4.1), b^0 is a Poisson rv with parameter $\rho = \lambda \overline{\sigma}$ and $\{\hat{\sigma}_j, j \in \mathbb{N}^\star\}$ are iid rv's, with common cumulative probability distribution G_1, independent of b^0.

Furthermore, the rv's $\{T_j, \sigma_j,\ j \in \mathbb{N}^\star\}$ are independent of the rv's $\{b^0, \hat{\sigma}_j,\ j \in \mathbb{N}^\star\}$. $\diamond$

The first term (resp. second term) in the r.h.s. of (4.1) describes the contribution to the number of customers in the system at times $t \in \mathbb{N}$ from those present just before time $t = 0$ (resp. from the new arrivals in $[0, t)$) The rv's b^0 and $\hat{\sigma_j}$ represent the number of busy servers in steady-state and the residual lifetime of the rv σ_j, respectively.

An appealing feature of the stationary version of the M/G/∞ process[1] is that its autocovariance is known in closed form. More precisely [10, p. 139]

$$r(k) = \overline{G_1}(k), \quad k \in \mathbb{N} \tag{4.2}$$

thereby showing that the M/G/∞ process exhibits positive correlations. In fact, the process $\{b^t, t \in \mathbb{N}\}$ can be shown [34] to be associated [16], in that for any $t \in \mathbb{N}$ and any pair of non–decreasing mappings $f, g : \mathbb{N}^{t+1} \to \mathbb{R}$, $\mathbf{E}[f(b^0, \cdots, b^t)\, g(b^0, \cdots, b^t)] \geq \mathbf{E}[f(b^0, \cdots, b^t)]\, \mathbf{E}[g(b^0, \cdots, b^t)]$, provided the expectations exist and are finite.

We may observe from (4.2) that the process $\{b^t,\ t \in \mathbb{N}\}$ will be long-range dependent if $\sum_{k \in \mathbb{N}} \overline{G_1}(k) = \infty$, which will occur, for instance, when G is Pareto with parameter $1 < \alpha < 2$.

Also worth pointing out is the fact that the process $\{b^t, t \in \mathbb{N}\}$ is *asymptotically second-order self-similar* with Hurst parameter $H = (3 - \alpha)/2$ when G is a Pareto-like distribution with parameter $1 < \alpha < 2$ [43, Appendix A].

[1] From now on we will only consider the stationary version $\{b^t, t \in \mathbb{N}\}$ of the M/G/∞ process, as defined in Lemma 4.1.

The rest of this section is devoted to the study of the queue length when the queue is fed by an M/G/∞ process. To this end, we first recall some standard results of queueing theory (e.g. see [4]).

As in Section 4 if the server (router) can transmit c packets per slot, then the queue length at time t satisfies the recursion

$$q_{t+1} = \max\{q_t + b_t - c, 0\}, \quad t \in \mathbb{N}$$

for some initial condition $q_0 = Q$.

Since the process $\{b_{t+k}, t \in \mathbb{N}\}$ converges in distribution to the stationary and ergodic process $\{b^t, t \in \mathbb{N}\}$ as $k \to \infty$ by Lemma 4.1, it is well-known that under the stability condition $\rho = \mathbf{E}[b^0] < c$, q_t converges in distribution to a proper rv q [4, Theorem 6, p. 12]. Here, the stationary sequence $\{b^t, t \in \mathbb{N}\}$ being also reversible for any distribution G (as the process of the number of busy servers in an M/G/∞ is reversible [21, Theorem 3.11]), the tail distribution of q is given by

$$\mathbf{P}(q > x) = \mathbf{P}\left(\sup_{t \in \mathbb{N}} \left(\sum_{s=0}^{t-1} b^s - ct\right) > x\right), \quad x \in \mathbb{N}. \tag{4.3}$$

Needless to say that the task of finding an explicit expression for the r.h.s. of (4.3) is difficult, not to say more. We will instead content ourselves with an asymptotic lower bound on the tail distribution of the queue length. This bound will already reveal some interesting properties of the buffer statistics in presence of bursty traffic. From now on we shall assume that the stability condition $\rho < c$ is satisfied.

Proposition 4.1 (Asymptotic lower bounds on the queue length)

Let $\rho < c$. For any activity period distribution G,

$$\liminf_{t \to \infty} \frac{\mathbf{P}(q > t)}{\overline{G_1}(t)^N} \geq L \tag{4.4}$$

with

$$N = \begin{cases} c - \rho + 2 & \text{if } c - \rho \text{ is an integer number} \\ \lceil c - \rho \rceil + 1 & \text{otherwise} \end{cases} \tag{4.5}$$

and

$$L = 1 - \sum_{k=0}^{N} \frac{\rho^k}{k!} e^{-\rho} > 0. \tag{4.6}$$

Proof. Define $\gamma(\epsilon) = c - \rho + 1 + \epsilon$ with $0 < \epsilon < \min\{\rho, 1 + c - \rho - \lceil c - \rho \rceil\}$. Clearly, $\lceil \gamma(\epsilon) \rceil = N$ where N is defined in (4.5). Also note that $\rho - \epsilon > 0$.

Let $A(t) = \sum_{s=0}^{t-1} b^s$ be the number of customers entering the queue in $[0, t)$. With (4.1) we find

$$A(t) = \sum_{s=0}^{t-1} a_s(t) \tag{4.7}$$

with

$$a_0(t) = \sum_{j=1}^{b^0} \min\left(\lceil \hat{\sigma}_j \rceil, t\right) \tag{4.8}$$

and

$$a_s(t) = \sum_{s-1 \le T_j < s} \sum_{i=s}^{t-1} I(\sigma_j > i - T_j), \quad s = 1, 2, \ldots, t-1. \tag{4.9}$$

Starting from (4.3) and using the definition of $A(t)$ we get

$$\begin{aligned} \mathbf{P}(q > t) &\geq \mathbf{P}\left(A(t) - ct \geq t\right) \\ &\geq \mathbf{P}\left(a_0(t) \geq \gamma(\epsilon)t,\ \sum_{s=1}^{t-1} a_s(t) > (\rho - \epsilon)t\right) \\ &= \mathbf{P}(a_0(t) \geq \gamma(\epsilon)t)\, \mathbf{P}\left(\sum_{s=1}^{t-1} a_s(t) > (\rho - \epsilon)t\right) \end{aligned} \tag{4.10}$$

where (4.10) follows from the independence of the rv's $a_0(t)$ and $\{a_s(t), s = 1, 2, \ldots, t-1\}$ (Lemma 4.1).

Consider the first factor in the r.h.s. of (4.10). Conditioning on b^0, using the independence of the rv's $\{\hat{\sigma}_j, j = 1, 2 \ldots\}$ and b^0 (Lemma 4.1), along with the identity $N = \lceil \gamma(\epsilon) \rceil$, gives

$$\begin{aligned} \mathbf{P}(a_0(t) \geq \gamma(\epsilon)t) &\geq \sum_{k=N}^{\infty} \mathbf{P}\left(\sum_{j=1}^{k} \min(\hat{\sigma}_j, t) \geq \gamma(\epsilon)t\right) \mathbf{P}(b^0 = k) \\ &\geq \sum_{k=N}^{\infty} \mathbf{P}\left(\hat{\sigma}_1 > t, \cdots, \hat{\sigma}_N > t\right) \mathbf{P}(b^0 = k) \\ &= \overline{G_1}(t)^N\, \mathbf{P}(b^0 \geq N) \quad \text{for } t = 1, 2, \ldots. \end{aligned} \tag{4.11}$$

On the other hand, the ergodicity of the sequence $\{b^t, t \in \mathbb{N}\}$ which implies that [41, Chapter V])

$$\lim_{t \to \infty} \frac{1}{t} \sum_{s=0}^{t-1} b^s = \mathbf{E}[b^0] = \rho \quad \text{a.s.},$$

combined with the identity $\sum_{s=0}^{t-1} b^s = a_0(t) + \sum_{s=1}^{t-1} a_s(t)$ (cf. (4.7)-(4.9)) yields

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

as we trivially have $\lim_{t\to\infty} a_0(t)/t = 0$ a.s.. Consequently,

$$\lim_{t\to\infty} \mathbf{P}\left(\sum_{s=1}^{t-1} a_s(t) > (\rho-\epsilon)t\right) = 1. \tag{4.12}$$

Dividing now both sides of (4.10) by $\overline{G_1}(t)^N$, taking the liminf as $t \to \infty$ and using (4.11) and (4.12), gives

$$\liminf_{t\to\infty} \frac{\mathbf{P}(q>t)}{\overline{G_1}(t)^N} \geq \mathbf{P}\left(b^0 \geq N\right) \tag{4.13}$$

which concludes the proof. ∎

The asymptotic lower bound (4.4) indicates that the tail distribution of the queue length does not decrease to zero faster than the integrated tail distribution of an activity period raised at the power $N \geq 2$. To be more concrete, consider the case when G is a Pareto distribution, namely, $\overline{G}(x) \sim c_1\, x^{-\alpha}$ $(x \to \infty)$ with $\alpha > 1$ (to ensure the existence of the first moment) and $c_1 > 0$ Hence,

$$\overline{G_1}(x) \sim c_2\, x^{-\alpha+1} \quad (x \to \infty) \tag{4.14}$$

with $c_2 = c_1/(\overline{\sigma}(\alpha-1))$. From (4.4) we get

$$\liminf_{x\to\infty} \frac{P(q>x)}{x^{(-\alpha+1)N}} \geq L\, c_2^{(\alpha-1)N} \tag{4.15}$$

In other words, the tail of the queue length inherits the heavy-tailed nature of the Pareto distribution of an activity period.

To derive (4.14) we have only assumed that $\alpha > 1$. In particular (4.14) will hold if $\alpha > 2$, a situation where the M/G/∞ process is neither asymptotically second-order self-similar nor long-range dependent. This shows that self-similarity and/or long-range dependence are not necessary to produce heavy-tailed queue lengths, a conclusion that can also be reached from an earlier result of Pakes [33] and Veravebeke [44].

Since Proposition 4.1 holds for any distribution G it holds, in particular, for moderate tail distributions such as the Log-normal and the Weibull distributions. Therefore, Proposition 4.1 tells us that the queue length will not be lighter than that of a moderate tail rv if G is itself moderate, a situation that is again very different from the situation observed under non-bursty traffic, as discussed in Section 3.

Figure 2 reports simulation results. It displays the mapping $x \to \log_{10} \mathbf{P}(q > x)$ for two different probability distributions G: $G(x) = 1 - \exp(-\mu x)$ (Exponential distribution with mean $\overline{\sigma}_{\rm E} = 1/\mu$) and $G(x) = 1 - (a/(x+a))^\alpha$ (Pareto distribution with mean $\overline{\sigma}_{\rm P} = a/(\alpha-1)$). Note from (4.2) that even when G is exponential the M/G/∞ process $\{b^t, t \in \mathbb{N}\}$ is *not* a renewal process since $r(k) = \exp(-\mu k) > 0$. We have chosen $\alpha = 1.5$ so that the M/G/∞ process is asymptotically second-order self-similar when G is Pareto, $\mu = 1.0$ and $a = \alpha - 1$ so that $\overline{\sigma} = \overline{\sigma}_{\rm E} = \overline{\sigma}_{\rm P} = 1$, $\lambda = 0.7$ and $c = 1$. Under these parameters the queue is stable as $\rho = 0.7 < c$. The results

have been obtained by simulation by using the modeling software QNAP2[2]. These curves illustrate the very different behavior of the buffer statistics for short-tailed (exponential) and heavy-tailed (Pareto) activity period distributions and show the key role played by the correlation structure of the input process on the performance. The accuracy of the asymptotic lower bound (4.4) has also been investigated in the case when G is a Pareto distribution. To this end, we have plotted in Figure 2 the mapping $t \to L\,\overline{G_1}(t)^N$ for $t \in \{0, 5, 10, 15, 20, 25, 30\}$. The results in Figure 2 seem to indicate that the asymptotic lower bound in (4.4) is fairly loose, thereby suggesting that $\mathbf{P}(q > x)$ may actually decrease much slower than $L\,\overline{G_1}(t)^N$, enhancing even more the impact of bursty traffic on queueing performance.

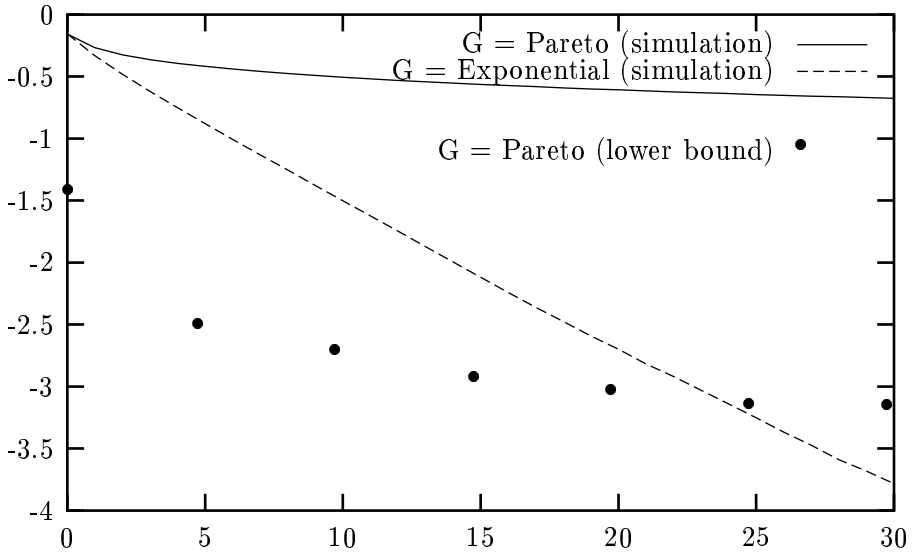


Figure 2: $x \to \log_{10} \mathbf{P}(q > x)$ for $G \in \{\text{Exponential, Pareto}\}$

We conclude this section with a few remarks.

Remarks 4.1

(1) When the distribution G_1 belongs to the class $\mathcal{D} \subset \mathcal{S}$ of *dominated-variation* distributions, then a tighter lower bound than (4.4) can be derived [29, Prop. 3.2]. A distribution $F \in \mathcal{D}$ if $\limsup_{x\to\infty} \overline{F}(x)/\overline{F}(2x) < \infty$. Pareto-like distributions belong to $\mathcal{D}$ but Log-normal and Weibull distributions do not. Note, however, that the bound in (4.4) is more "versatile" than all other bounds reported to date since it holds for *any* activity period distribution G.

(2) The lower bound (4.4) can be complemented with an upper bound [29] whenever G and G_1 are both subexponential distributions. This occurs, for instance, if $G_1 \in \mathcal{S}$ if G either Pareto,

[2]QNAP2 is a trademark by INRIA. QNAP2 is marketed by SIMULOG – http://www.simulog.fr/US/welcome.html

Log-normal or Weibull. In this case the upper bound reads

$$\limsup_{x\to\infty} \frac{\mathbf{P}(q > x)}{\overline{G_1}(x)} \leq \rho + \frac{\rho}{c-\rho}. \tag{4.16}$$

(3) When $c - \rho < 1$, it was shown in [29] that

$$\lim_{x\to\infty} \frac{1}{\log x} \log \mathbf{P}(q > x) = -\alpha + 1$$

when G is Pareto, and

$$\lim_{x\to\infty} \frac{1}{(\log x)^2} \log \mathbf{P}(q > x) = -\frac{1}{2\delta^2}$$

when G is Log-normal. Again, these limiting results contrast with the corresponding result (3.3) found for non-bursty traffic.

(4) The condition $c - \rho < 1$ appearing in item (3) above identifies a situation when the available capacity c is not sufficient to process all customers generated by a single session, in addition to the average traffic ρ. This is an example of a situation where a single session may dictate the performance.

(5) During the revision of this paper we became aware of two recent works [26] and [39] (for the continuous-time version of the model considered here) in which the authors were able to derive the exact asymptotics for the tail of the queue length distribution when the input is a M/G/∞ process with Pareto-like activity period distribution and under the condition $\rho < c$. As expected, their results show that the tail of q is heavier than the tail of σ.

5 Concluding remarks

In this paper we have shown by using a simple, yet pertinent traffic model, the impact of bursty traffic on queueing performance. However, extrapolating these preliminary results to real networks requires some care since several network characteristics do not appear in our model. These characteristics include closed-loop flow control mechanisms (TCP in the Internet) that aim at regulating traffic at the sender and finite buffers in routers. Buffers are necessarily finite but in practice they may even be fairly small so as to prevent long delays from building up. The conjunction of closed-loop control schemes and finite buffering should "in principle" break up correlations, thereby advocating the use of traffic models with correlations only up to some finite lags (proportional to the buffer size?) as it is the case in ... Markovian models! The debate is therefore still wide open between pro-LRD and pro-Markovian traffic models and much work is needed, including further measurement campaigns, to reach a better understanding of the impact of the existence of correlations at multiple time scales on the QoS delivered by networks to the end users.

Acknowledgements: The author thanks the anonymous referees for their comments and suggestions which helped improving the presentation of this paper.

Miscellaneous: *This paper contains material presented at the 19th French-Belgian Workshop of Statisticians on "Limit Theorems and Long-Range Dependances in Statistics", which was held in Marseille, France, in the Fall of 1998. Proposition 4.1 is inspired by a collaborative work [29] with Z. Liu (INRIA), D. Towsley (Univ. Massachusetts, MA) and Z.-L. Zhang (Univ. Minnesota, MN).*

References

[1] R. G. Addie, M. Zukerman and T. Neame, "Fractal traffic: Measurements, modeling and performance evaluation," *Proc. of the IEEE Infocom'95 Conf.*, Boston, MA, Apr. 4-6, 1995, 977-984.

[2] R. Agrawal, A. M. Makowski and P. Nain, "On a reduced load equivalence for fluid queues under subexponentiality," *Queueing Systems and Its Applications (QUESTA)*, Vol. 33, No. 1-3, pp. 5-41, 1999.

[3] J. Beran, *Statistics for Long-Memory Processes*. Chapman and Hall, New York, 1994.

[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*, Vol. 27&28, pp. 699-712, 1996.

[6] O. J. Boxma and V. Dumas, "Fluid queues with long-tailed activity period distributions," *Computer Communications*, Vol. 21, pp. 1509-1529, 1998.

[7] F. Brichet, J. W. Roberts, A. Simonian and D. Veitch, "Heavy traffic analysis of a storage model with long range dependent on/off sources," *Queueing Systems*, Vol. 23, pp. 197-215, 1996.

[8] C.-S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks", *IEEE Trans. Aut. Contr.*, Vol. 39, No. 5, pp. 913–931, May 1994.

[9] G. L. Choudhury and W. Whitt, "Long-tail buffer-content distributions in broadband networks," *Performance Evaluation*, Vol. 30, pp. 177-190, 1997.

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

[11] M. Crovella and A Bestavros, "Self-similarity in world wide Web traffic: evidence and possible causes," *Proc. 1996 ACM Sigmetrics Int. Conf. on Measurements and Modeling of Comput. Syst.*, May 1996.

[12] N. G. Duffield, "Exponential bounds for queues with Markovian arrivals", *Queueing Systems*, Vol. 17, pp. 413-430, 1994.

[13] 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.

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

[15] A. Erramilli, O. Narayan and W. Willinger, "Experimental queueing analysis with long-range dependence packet traffic," *IEEE/ACM Trans. on Networking*, Vol. 4, pp. 209-223, 1996.

[16] J. D. Esary, F. Proschan and D. W. Walkup, "Association of random variables with applications", *Annals of Math. Stat.*, Vol. 38, pp. 1466-1474, 1967.

[17] D. Heath, S. Resnick and G. Samorodnitsky, "Patterns of buffer overflow in a class of queues with long memory in the input stream," *Annals of Applied Prob.*, Vol. 7, pp. 1021-1057, 1997.

[18] P. W. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," *Studies in Appl. Prob.*, J. Galambos and J. Gani Eds, pp. 131–156, 1994.

[19] P. Jacquet, "Long term dependences and heavy tails in traffic and queues generated by memoryless on/off sources in series," INRIA Research Report No. 3516, Oct. 1998.

[20] P. R. Jelenkovic and A. A. Lazar, "Asymptotic results for multiplexing on-off sources with subexponential on periods," *Adv. in Appl. Prob.*, Vol. 31, No. 2, Jun. 1999.

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

[22] J. F. C. Kingman, "Inequalities in the theory of queues," *J. Roy. Stat. Soc.*, Series B, Vol. 32, pp. 102–110, 1970.

[23] L. Kleinrock, *Queueing Systems, Vol. I.* J. Wiley & Sons, New York, 1975.

[24] M. M. Krunz and A. M. Makowski, "Modeling video traffic using M/G/∞ input processes: A compromise between Markovian and LRD models," *IEEE J. of Selected Areas in Comm.*, Vol. 16, No. 5, pp. 733-748, Jun. 1998.

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

[26] N. Likhanov and R. Mazumdar, "Cell loss asymptotics in buffers fed by heterogeneous long-tailed sources." *Proc. of the IEEE INFOCOM Conf.*, Tel-Aviv, Israel, Mar. 26-30, 2000

[27] N. Likhanov, B. Tsybakov and 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.

[28] Z. Liu, P. Nain and D. Towsley "Exponential bounds with application to call admission," *J. of the ACM*, Vol. 44, No. 3, pp. 366-394, May 1997.

[29] Z. Liu, P. Nain, D. Towsley and Z.-L. Zhang, "Asymptotic behavior of a multiplexer fed by a long-range dependent process," *J. of Appl. Prob.*, Vol. 36, No. 1, pp. 105-118, Mar. 1999.

[30] B. B. Mandelbrot and J. W. van Ness, "Fractional Brownian motions, fractional noises and applications", *SIAM Review*, Vol. 10, No. 4, pp. 422-437, Oct. 1968.

[31] I. Norros, "A storage model with self-similar input," *Queueing Systems and Its Applications (QUESTA)*, Vol. 16, pp. 387-396, 1994.

[32] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks," *IEEE J. on Sel. Areas in Comm.*, Vol 13, No. 6, pp. 953-962, Aug. 1995.

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

[34] M. Parulekar *Buffer Engineering for $M|G|\infty$ Traffic Models: Analysis and Simulations*, Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, Dec. 1999.

[35] M. Parulekar and A. M. Makowski, "Tail probabilities for M/G/∞ input processes (I): Preliminary asymptotics," *Queueing Systems and Its Applications (QUESTA)*, Vol. 27, pp. 271-296, 1997.

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

[37] V. Paxson and S. Floyd, "Wide-area traffic: the failure of Poisson modeling," *Proc. of the ACM Sigcomm'94 Conf.*, London, UK, pp. 257-268, 1994.

[38] S. Resnick and G. Samorodnitsky, "Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues," *Queueing Systems and Its Applications (QUESTA)*, Vol. 33, No. 1-3, pp. 43-71, 1999.

[39] S. Resnick and G. Samorodnitsky, "Steady-state distribution of the buffer content for M/G/∞ input fluid queues," *Bernoulli*, Vol. 7, pp. 191-210, 2001.

[40] T. Rolski, S. Schlegel and V. Schmidt, "Asymptotics of Palm-stationary buffer content distributions in fluid queues," *Adv. in Appl. Prob.*, Vol. 31, pp. 235-253, 1999.

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

[42] L. Takács, *Theory of Queues*. Oxford University Press, New York, 1962.

[43] B. Tsybakov and N. D. Georganas, "On self-similar traffic in ATM queues: Definitions, overflow probability and cell delay distribution," *IEEE/ACM Trans. on Networking*, Vol. 5, No.3, pp. 397-409, Jun. 1997.

[44] N. Veraverbeke, "Asymptotic behaviour of Wiener-Hopf factors of a random walk", *Stoch. Proc. and their Applications*, Vol. 5, pp. 27-37, 1977.

[45] W. Willinger, M. S. Taqqu and R. Sherman, "Proof of a fundamental result in self-similar traffic modeling," *Computer Communication Review*, Vol. 27, No. 2, pp. 5-23, Apr. 1997.