Multiplexing On-Off Sources with Subexponential On Periods: Part II

Predrag R. Jelenković[a] and Aurel A. Lazar[b]

[a] Bell Laboratories, Innovations for Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ 07974

[b]Department of Electrical Engineering, and Center for Telecommunications Research
Columbia University, New York, NY 10027

We consider an aggregate arrival process A^N obtained by multiplexing N On-Off sources with exponential Off periods of rate λ and generally distributed On periods τ^{on}. When N goes to infinity, with $\lambda N \to \Lambda$, A^N approaches an $M/G/\infty$ type process. For a fluid queue with the limiting M/G/∞ arrivals A_t^∞, regularly varying On periods with noninteger exponent, and capacity c, we obtain a *precise asymptotic behavior* of the queue length random variable Q_t^P observed at the beginning of the arrival process activity periods

$$\mathbb{P}[Q_t^P > x] \sim \Lambda \frac{r+\rho-c}{c-\rho} \int_{x/(r+\rho-c)}^{\infty} \mathbb{P}[\tau^{on} > u] du \quad x \to \infty,$$

where $\rho = \mathbb{E}A_t^\infty < c$; r $(c \le r)$ is the rate at which the fluid is arriving during an On period. (In particular, when $\mathbb{P}[\tau^{on} > x] \sim x^{-\alpha}, 1 < \alpha < 2$, the above formula applies to the so-called long-range dependent On-Off sources.) Based on this asymptotic result and the results from a companion paper we suggest a computationally efficient approximation for the case of finitely many long-tailed On-Off sources. The accuracy of this approximation is verified with extensive simulation experiments.

1. Introduction

The problem of multiplexing On-Off sources arises frequently as the basic model of contention in multimedia communication systems (as well as in some storage systems). The most basic mathematical model for this problem is an infinite buffer queue loaded with (multiplexed) On-Off arrivals; for ATM based communication systems the main performance measure (Quality of Service parameter) is the buffer overflow probability distribution.

Quantitative analysis of the buffer overflow probability distribution for this queueing problem dates back to [37,13]. In [13], Cohen obtained a complete Laplace transform solution to this problem! (More recently, he revisited the problem in [14].) However, inverting the Laplace transform is usually a very tedious process. Hence, it is necessary to investigate computationally tractable exact and approximate solution techniques. For Markovian (fluid) On-Off sources, a thorough investigation of this problem was done in [2]. Many other results for multiplexing finite Markovian On-Off sources followed. These led to the Equivalent Bandwidth theory for finite Markovian (or, in general, exponentially bounded) arrival processes; extensive references can be found, for example, in [17,18,16].

Recently, statistical analysis has increasingly shown that the traffic streams in modern broadband networks exhibit long-tailed characteristics. [29] pointed out first the existence of long-tailed statistics in Ethernet traffic. These statistical results have stimulated research in queueing analysis under the heavy tailed (non Cramér) assumptions. Queueing analysis with self-similar long-range dependent arrival processes appear in [32,16,30,35,38, 36]. Recently, long-tailed characteristics of the scene length distribution of MPEG video streams were explored in [25,20,26,22].

Parallel to the modeling approach through self-similar long-range dependent processes, a more analytically tractable approach using fluid renewal type models, in which inter-renewal times are long-tailed, has been explored in [1,19]. Queueing results in these two papers rely on the classical result by Pakes [33] on the subexponential (long-tailed) asymptotics of the waiting time distribution in a GI/GI/1 queue (or the earlier work of Cohen [12] which considered a regularly varying GI/GI/1 queue).

Recently, the result of Pakes has been generalized to a Markov modulated setting [4,24]. In [4] the subexponential asymptotics of a Markov modulated M/G/1 queue was investigated. Work in [24] further generalized these results to Markov modulated G/G/1 queues. In the same paper it was shown that a subexponential GI/GI/1 queue is invariant under Markov modulation. In other words, a subexponential Markov modulated G/G/1 queue has the same asymptotics as the corresponding GI/GI/1 queue. These results made possible the analysis of a subexponential semi-Markov fluid queue ([24]).

The analysis of a fluid queue in which more than one long-tailed source is multiplexed appears to be a very difficult problem. This is due to the fact that the renewal structure of an aggregate arrival process may be very complex, although the appearance of each individual source may be truly innocuous (like an On-Off source). The complex autocorrelation structure of the aggregate source obtained by multiplexing long-tailed On-Off sources was examined in [19]. General bounds for multiplexing long-tailed fluid processes was obtained in [9]. In [7] a limiting case of an infinite number of On-Off sources with regularly varying On distribution was investigated. In the same paper a case of two sources, in which one source had regularly varying On periods, and the other had exponential On periods, was solved. The literature does not explicitly give precise asymptotic results for the case of multiplexing two or more long-tailed sources.

More precise asymptotic results for multiplexing long-tailed On-Off sources have recently been obtained in [23] (see also [21]). From an engineering standpoint, this paper advances two important results. The first result intuitively says that when a source with subexponential characteristics (e.g., MPEG video) is multiplexed with a source that has exponential characteristics (e.g., voice), the contribution to the large buffer asymptotics of the exponential sources is reflected only through their mean values. This result suggests that (under the appropriate conditions) for admission control of both VBR video and voice streams, the voice streams need to be characterized only by their means. The second result is a general asymptotic lower bound on the buffer overflow probabilities for multiplexing a large number of On-Off sources.

The main result derived in this paper proves that the lower bound obtained in [23] is exact. Based on this asymptotic result, we suggest an approximation for the buffer overflow probabilities. We verify the accuracy of the approximation with extensive simulation experiments. Besides accuracy, it is of a special importance for engineering the MUX that this approximation has basically *negligible computational complexity!* To the best of our knowledge, this is the only result in the literature (of comparable computational complexity) that is both proven theoretically and demonstrated experimentally as a good approximation for the buffer overflow probabilities with multiplexed long-tailed arrivals.

The rest of the paper is organized as follows. Section 2 contains necessary definitions and examples of long-tailed and subexponential distributions. In Section 3, we examine the aggregate arrival process obtained by multiplexing a large number ($N \to \infty$) independent and identical On-Off sources with regularly varying (with noninteger exponent) On periods. Using Karamata's theory, we obtain a precise asymptotic behavior of the server overflow distribution during the arrival process activity period. In Section 4, using these asymptotic relations, we derive a precise fluid queue asymptotics with multiplexed long-tailed On-Off arrivals. In the same section, based on queueing theoretic results, we suggest a computationally efficient approximation for multiplexing a finite number of subexponential On-Off sources. The paper is concluded in Section 5.

2. Long Tailed and Subexponential Distributions

This section contains the necessary definitions of long-tailed and subexponential distributions. Since it is not obvious what distributions belong to these classes, immediately from the definitions, we list several well known examples. For more details on long-tailed and subexponential distributions the reader is referred to [28].

Definition 1 *A distribution function F on $[0, \infty)$ is called* long-tailed *($F \in \mathcal{L}$) if*

$$\lim_{x\to\infty} \frac{1 - F(x-y)}{1 - F(x)} = 1, \quad y \in \mathbb{R}. \tag{1}$$

Definition 2 *A distribution function F on $[0, \infty)$ is called* subexponential *($F \in \mathcal{S}$) if*

$$\lim_{x\to\infty} \frac{1 - F^{*2}(x)}{1 - F(x)} = 2, \tag{2}$$

where F^{*2} denotes the 2-nd convolution of F with itself, i.e., $F^{*2}(x) = \int_{[0,\infty)} F(x - y)F(dy)$.

The class of subexponential distributions was first introduced by Chistakov [8]. The definition is motivated by the simplification of the asymptotic analysis of the convolution tails. Some well known families of distribution functions in $\mathcal{S}$ (and $\mathcal{L}$) are: regularly varying, lognormal, Weibull ($e^{-x^\alpha}, 0 < \alpha < 1$), Benktander Type I, II, [28].

In this paper functions of Regular ($\mathcal{R}_{-\alpha}$) will be of our main interest. This functions were invented by Karamata [27] (the main reference is [6]).

Definition 3 *$F \in \mathcal{R}_{-\alpha}$ if it is given by*

$$1 - F(x) = \frac{l(x)}{x^\alpha}, \qquad \alpha \geq 0,$$

where $l(x) : \mathbb{R}_+ \to \mathbb{R}_+$ is a function of slow variation, i.e., $\lim_{x\to\infty} l(\delta x)/l(x) = 1, \forall \delta > 1$.

(Karamata's motivation was to derive Tauberian/Abelian theorems, i.e., to establish *relationship between the asymptotic behavior of a function at infinity and the asymptotic behavior of its Laplace transform at zero.*)

3. Analysis of the Aggregate Arrival Process

In this section we asymptotically characterize the aggregate arrival process functionals that are relevant to further our queueing investigation. Our main result is given in Subsection 3.1. There, we derive the asymptotic behavior of the distribution of the server overflow distribution during the arrival process activity period. This is achieved by applying Karamata's theory. In Section 4, these results will be used to obtain the already advertised asymptotic queueing results.

More formally, consider a sequence of i.i.d. random variables $\{\tau_n^{off}, \tau_n^{on}, n \geq 0\}$, $\tau_0^{off} = \tau_0^{on} = 0$. Define a point process $T_n^{off} \overset{\text{def}}{=} \sum_{i=0}^{n}(\tau_i^{off} + \tau_i^{on}), n \geq 0$; this process will be interpreted as representing the beginnings of Off periods in an On-Off source. Further, define an On-Off source a_t with rate r, as

$$a_t = r \ \text{ if } \ T_n^{off} - \tau_n^{on} \leq t < T_n^{off}, \ \ n \geq 1,$$

and $a_t = 0$, otherwise. For the rest of the paper, unless otherwise specified, we will assume that τ_n^{off} is exponentially distributed with parameter λ, i.e., $\mathbb{P}[\tau_n^{off} > t] = e^{-\lambda t}, t \geq 0$. Also, τ_n^{on} is assumed to have a finite mean. Steady state probabilities of this process are given as $\pi_0 = \lim_{t\to\infty} \mathbb{P}[a_t = 0] = 1/(1 + \lambda \mathbb{E}\tau^{on}) = 1 - \pi_1$, where $\pi_1 = \lim_{t\to\infty} \mathbb{P}[a_t = r]$. Let $A^N = \sum_1^N a^i$, be an aggregate arrival process obtained by multiplexing N independent and identical On-Off sources $a^i, 1 \leq i \leq N$.

Infinite number of sources. Now, we will investigate the Poisson type limit of the aggregate arrival process. Let $T_n, n \geq 0, T_0 = 0$, be a Poisson process with rate Λ. Define $A_t^\infty = \sum_{n=1}^\infty r 1(T_n \leq t < T_n + \tau_n^{on}), r > 0$. Then the following theorem holds.

Theorem 1 *If $\mathbb{E}\tau_n^{on} < \infty$, and $\lambda N \to \Lambda$ as $N \to \infty$, then*

$$A_t^N \overset{d}{\Rightarrow} A_t^\infty, \ \text{ as } \ N \to \infty, \tag{3}$$

where $\overset{d}{\Rightarrow}$ symbolizes convergence in distribution.

Proof: It is enough to prove that the beginnings of the On periods in the process A_t^N converge to a Poisson process with rate Λ. This follows from a classical result on multiplexing a large number of renewal processes [15,10]. ◇

Lemma 1 *The transient probability of the arrival process A_t^∞ being silent is given by*

$$\mathbb{P}[A_t^\infty = 0] = e^{-\Lambda \int_0^t \mathbb{P}[\tau^{on} > u] du}. \tag{4}$$

Furthermore, if $\mathbb{E}\tau^{on} < \infty$, then $\lim_{t\to\infty} \mathbb{P}[A_t^\infty = 0] = e^{-\Lambda \mathbb{E}\tau^{on}}$.

Proof: Follows from Theorem 2.2 in [13]. ◇

Remark: Observe that $\mathbb{P}[A_t = 0] = \mathbb{P}[V_t = 0]$, where V_t is the workload process of an M/GI/∞ queue in which $V_0 = 0$, the customer service requirement has the same distribution as τ^{on}, and the arrival rate is Λ. (For recent asymptotic results on M/G/∞ processes see [35].)

3.1. Total Server Overflow During the Activity Period

Let $B_n, n \geq 1$, be a sequence of random variables representing the total amount of fluid that is brought to the system during the nth activity period, i.e., $B_n = \int_{t_n^b}^{t_n^e} A_t^\infty dt$, where t_n^b, t_n^e, represent the beginning, and the end of the nth activity period, respectively. Further, define $D_n^c \overset{\text{def}}{=} B_n - cI_n^{on}, 0 < c \leq r$; note that $D_n \equiv D_n^c$ is a non-negative random variable. If we imagine that A_t^∞ represents the rate at which the fluid is arriving to a fluid queue, and that c is the constant rate at which the queue drains, then D_n represents the queue increment (server overflow) during the nth activity period. Therefore, in order to solve the queueing asymptotics, we first have to understand the asymptotic behavior of D_n. Unfortunately, this is a much more difficult task than the investigation of the asymptotic behavior of the activity period that was done in [23]. For that reason we are forced to work under much more restrictive assumptions of distribution functions of regular variation. The method of proof, for the following result, will be through Karamata's Tauberian/Abelian theorems.

Theorem 2 *Assume that the distribution of On periods is regularly varying* $\mathbb{P}[\tau^{on} > x] = l(x)/x^\alpha, \alpha > 1$, *where* α *is noninteger. Then,*

$$\mathbb{P}[D_n^c > x] \sim e^{\Lambda \mathbb{E}\tau^{on}} \mathbb{P}\left[\tau^{on} > \frac{x}{r + r\Lambda\mathbb{E}\tau^{on} - c}\right] \quad as \quad x \to \infty. \tag{5}$$

Proof: Given in [21]. ◇

Next, consider a stationary version of the arrival process $A_t^{\infty,s} = \sum_{-\infty<n<\infty} r1(T_n \leq t < T_n + \tau_n^{on})$, where T_n is a stationary Poisson process with rate Λ. Given that at time $t = 0$, the arrival process is active ($A_t > 0$), denote with $D_{(0)}^c$ the total queue increment since the beginning of the last activity period till time zero, i.e., $D_{(0)}^c = \int_{t_0^b}^0 (A_t^\infty - c)dt, 0 < c \leq r$, where t_0^b represent the beginning of the activity period that is still active at $t = 0$.

Now, by Theorem 4.3, pp. 64, [3], it follows that process $\{T_n + \tau_n^{on}, -\infty < n < \infty\}$ is also a stationary Poisson process with the same rate Λ. Therefore, process $A_t^{\infty,s}$ is reversible. This implies that $D_{(0)}^c$ is equal in distribution to $\int_0^{t_0^e} (A_t^\infty - c)dt$, where t_0^e represents the end of the activity period that is active at $t = 0$. For simplicity we will refer to both of these variables with $D_{(0)}^c$.

Conjecture 1 *Assume that the distribution of On periods is regularly varying* $\mathbb{P}[\tau^{on} > x] = l(x)/x^\alpha, \alpha > 1$, *where* α *is noninteger. Then,*

$$\mathbb{P}[D_{(0)}^c > x] \sim \frac{\Lambda e^{\Lambda\mathbb{E}\tau^{on}}}{e^{\Lambda\mathbb{E}\tau^{on}} - 1} \int_{x/(r+r\Lambda\mathbb{E}\tau^{on}-c)}^{\infty} \mathbb{P}\left[\tau^{on} > u\right] du \quad as \quad x \to \infty. \tag{6}$$

Heuristics: Due to space limitation this is given in [23]. ◇

4. Queueing Analysis

We start this section with a classical result on subexponential asymptotics of a GI/GI/1 queue. The result was obtained by Pakes 1975 (see also Veraverbeke for the random walk approach to this problem). For extensions of this result to Markov-modulated M/G/1 queues see [4], and to Markov-modulated G/G/1 queues see [24].

Let $X_n, n \geq 0$, be a sequence of i.i.d. random variables that are driving a queueing process (Lindley's recursion)

$$Q_{n+1} = (Q_n + X_n)^+, \ \ n \geq 0, \tag{7}$$

where $q^+ = \max(0, q)$. According to the classical result of Loynes' [31] under the stability condition $\mathbb{E}X_n < 0$ this recursion admits an unique stationary solution, and for all initial conditions $\mathbb{P}[Q_n \leq x]$ converges to the stationary distribution $\mathbb{P}[Q \leq x]$. For the rest of this paper we will assume that all the queueing systems under consideration are in their stationary regimes. Let G and $G_1(x) \overset{\text{def}}{=} 1/\mathbb{E}X_n \ \int_x^\infty \mathbb{P}[X_n > u]du$ represent the distribution and its integrated tail distribution for X_n, respectively.

Theorem 3 *If $G \in \mathcal{L}$, $G_1 \in \mathcal{S}$, and $\mathbb{E}X_n < 0$, then*

$$\mathbb{P}[Q_n > x] \sim \frac{1}{-\mathbb{E}X_n} \int_t^\infty \mathbb{P}[X_n > u]du \ \ \textit{as} \ \ t \to \infty.$$

Remark: In [24] it was shown that exactly the same asymptotics holds for Markov-modulated G/G/1 queues (equivalently random walks).

The rest of this section is organized in the following four subsections. In Subsection 4.1 we give some general results on the fluid flow queue. In order to develop some intuition about the behavior of the fluid flow queue with subexponential arrivals, we present in Subsection 4.2 a complete solution to a simple On-Off single server queue. Our main queueing theoretical results are presented in Subsection 4.3. A practical approximation technique based on the developed theoretical results is tested on simulation experiments in Subsection 4.4.

4.1. Fluid Queue: Preliminaries

The physical interpretation for a fluid queue is that at any moment of time t, fluid is arriving to the system with rate a_t, and is leaving the system with rate c_t. We term a_t, and c_t, to be the arrival, and service processes, respectively. Then, the evolution of the amount of fluid Q_t (also called queue length) evolves according to

$$dQ_t = (a_t - c_t)dt \ \ \text{if} \ \ Q_t > 0, \ \ \text{or} \ \ a_t > c_t, \tag{8}$$

and $dQ_t = 0$, otherwise. It is not very difficult to see that, starting from $Q_0 = 0$, the solution $Q_t, t \geq 0$, to (8) is given by

$$Q_t = \sup_{0 \leq u \leq t} \int_u^t (a_u - c_u)du. \tag{9}$$

And, if a_t, and c_t are stationary, Q_t is equal in distribution to

$$\mathbb{P}[Q_t \leq x] = \mathbb{P}[\sup_{0 \leq u \leq t} W_u \leq x],$$

where $W_t \overset{\text{def}}{=} \int_{-t}^0 (a_u - c_u)du, t \geq 0$. Now, whenever the stability condition $\mathbb{E}a_t < \mathbb{E}c_t$ is satisfied (by Birkhoff's Strong Law of Large Numbers), $\mathbb{P}[Q_t \leq x]$ converges to a proper probability distribution, i.e.,

$$\mathbb{P}[Q \leq x] \overset{\text{def}}{=} \lim_{t \to \infty} \mathbb{P}[Q_t \leq x] = \mathbb{P}[\sup_{0 \leq u < \infty} W_u \leq x].$$

4.2. Fluid Queue with a Single On-Off Source

Consider a fluid queue with capacity c and an On-Off arrival source with On arrival rate r. In this section we assume that Off periods are also general (not necessarily exponential). Then, if we observe the queue at the beginning of On periods, the queue length Q_n^P evolves as follows (P stands for Palm probability [5]).

$$Q_{n+1}^P = (Q_n^P + (r-c)\tau_n^{on} - c\tau_n^{off})^+, \quad n \geq 0. \tag{10}$$

Recall that F and F_1 denote the distribution and the integrated tail distribution of τ^{on}.

Theorem 4 *If $r > c$, $(r-c)\mathbb{E}\tau_{on} < c\mathbb{E}\tau_{off}$, $F \in \mathcal{L}$, and $F_1 \in \mathcal{S}$, then*

$$\mathbb{P}[Q_n^P > x] \sim \frac{r-c}{c\mathbb{E}\tau_{off} - (r-c)\mathbb{E}\tau_{on}} \int_{x/(r-c)}^{\infty} \mathbb{P}[\tau^{on} > u]du \;\; as \;\; x \to \infty. \tag{11}$$

Proof: Given in [21]. ◇

The relationship between the palm probabilities and the time average probabilities is presented in the following theorem.

Theorem 5 *If $r > c$, $(r-c)\mathbb{E}\tau_{on} < c\mathbb{E}\tau_{off}$, $F \in \mathcal{L}$ and $F_1 \in \mathcal{S}$, then*

$$\mathbb{P}[Q_t > x] \;\sim\; \mathbb{P}[Q^P > x] + \frac{1}{\mathbb{E}\tau^{off} + \mathbb{E}\tau^{on}} \int_{x/(r-c)}^{\infty} \mathbb{P}[\tau^{on} > u]du \tag{12}$$

$$\sim\; K \int_{x/(r-c)}^{\infty} \mathbb{P}[\tau^{on} > u]du \;\; \text{as} \;\; x \to \infty, \tag{13}$$

where

$$K = \frac{r-c}{c\mathbb{E}\tau_{off} - (r-c)\mathbb{E}\tau_{on}} + \frac{1}{\mathbb{E}\tau^{off} + \mathbb{E}\tau^{on}}. \tag{14}$$

Remarks: (i) This theorem improves on known results in [36,9] which were obtained under the assumptions of τ^{on} being regularly varying; (ii) The same proof can be carried out to establish the relationship between the Palm and time averages in much more general settings like semi-Markov fluid queues.
Proof: Given in [21]. ◇

4.3. Subexponential M/G/∞ Arrival Process

In this section we present our main queueing results. First we develop a general asymptotic queue distribution lower bound. Then, under more restrictive assumptions, we prove that this bound is precise.

4.3.1. Lower Bound

For the lower bound we need the following definition.

Definition 4 *A distribution function F is* intermediate regular varying *$F \in \mathcal{IR}$ if*

$$\lim_{\delta \downarrow 1} \liminf_{t \to \infty} \frac{\bar{F}(\delta t)}{\bar{F}(t)}.$$

Remark: For recent results on distributions of intermediate regular variation we refer the reader to [11]. Some basic properties of $\mathcal{IR}$ are: $\mathcal{R} \subset \mathcal{IR} \subset \mathcal{S}$.

Here, we obtain a tight lower bound for the fluid queue asymptotics with M/G/∞ arrivals. For this fluid queue we denote its queue content process as Q_t^∞.

Theorem 6 *Let $\rho \overset{\text{def}}{=} \mathbb{E}A_t^{s,\infty} = \Lambda r \mathbb{E}\tau^{on} < c$. If $r(1 + \Lambda\mathbb{E}\tau^{on}) > c$, and $\tau^{on} \in \mathcal{IR}$, then*

$$\liminf_{x\to\infty} \frac{\mathbb{P}[Q_t^\infty > x]}{\int_{x/(r+\rho-c)}^\infty \mathbb{P}[\tau^{on} > u]du} \geq \frac{\Lambda r}{c-\rho}.$$

Proof: Given in [21]. ◇

4.3.2. Precise Queue Asymptotics

Let $Q_n^{P,\infty}$ be the queue size observed at the beginning of the nth activity period of the M/G/∞ arrival process.

Theorem 7 *Let $\rho = \mathbb{E}A_t^{s,\infty} = \Lambda r \mathbb{E}\tau^{on} < c$. If $c \leq r$, and τ^{on} is regularly varying with noninteger exponent $\alpha > 1$, then*

$$\lim_{x\to\infty} \frac{\mathbb{P}[Q_t^{P,\infty} > x]}{\int_{x/(\rho+r-c)}^\infty \mathbb{P}[\tau^{on} > u]du} = \Lambda\left(\frac{r}{c-\rho} - 1\right).$$

Proof: Let I_n^{off} be the length of the nth off period in the arrival process $A_t^{s,\infty}$. Then, the proof follows directly from Theorem 3, and Theorem 2, by taking $X_n \overset{\text{def}}{=} D_n^c - cI_n^{off}$, observing that $\mathbb{P}[X_n > x] \sim \mathbb{P}[D_n^c > x]$, as $x \to \infty$, and $\mathbb{E}X_n = (\Lambda r\mathbb{E}\tau - c)/\Lambda e^{\Lambda\mathbb{E}\tau}$. ◇

Theorem 8 *Let $\rho = \mathbb{E}A_t^{s,\infty} = \Lambda r \mathbb{E}\tau^{on} < c$. If Conjecture 1 holds, $c \leq r$, and τ^{on} is regularly varying with noninteger exponent $\alpha > 1$, then*

$$\lim_{x\to\infty} \frac{\mathbb{P}[Q_t^\infty > x]}{\int_{x/(\rho+r-c)}^\infty \mathbb{P}[\tau^{on} > u]du} = \Lambda\frac{r}{c-\rho}.$$

Proof: This theorem follows from Theorem 8, Conjecture 1, and exactly the same arguments as in the proof of Theorem 5. We skip the details. ◇

Remark: The asymptotic result in this theorem is the same as the lower bound obtained in Theorem 6.

4.4. Finite Number of Subexponential On-Off Sources: M/G/∞ Approximation

Based on Theorems 6, 7, and 8, we suggest that the queueing probabilities obtained by multiplexing N long-tailed On-Off sources $a_t^i, 1 \leq i \leq N$, are approximated as

$$\mathbb{P}[Q_t^N > x] \approx \frac{\Lambda^N r}{c^N} \int_{x/(r-c^N)}^\infty \mathbb{P}[\tau^{on} > u]du, \qquad (15)$$

where $c^N \overset{\text{def}}{=} c - N\mathbb{E}a_t^i$, and $\Lambda^N \overset{\text{def}}{=} N\mathbb{E}a_t^i/(r\mathbb{E}\tau^{on})$. We term this approximation an M/G/∞ approximation. This approximation is to be used when the queue is stable and $r + (N-1)\mathbb{E}a_t^i > c$ *is satisfied.*

For simulation purposes we consider a discrete time "fluid" queue. Correspondingly, we replace exponential Off periods, with geometrically distributed random variables $\mathbb{P}[\tau^{off} = t] = p(1-p)^{t-1}, t = 1, 2, 3, \ldots$. For On periods we consider the Pareto family $\mathbb{P}[\tau^{on} \geq t] = 1/t^{\alpha}, t = 1, 2, \ldots, \alpha > 0$. Here, for the discrete Pareto case we use

$$\mathbb{P}[Q_t^N = x] \approx \frac{\Lambda^N r}{c^N}(r - c^N)^{\alpha-1} x^{-\alpha}, \tag{16}$$

where c^N, and Λ^N are as defined earlier.

The efficacy of approximation (16) is tested on numerous simulation experiments. For all experiments we fix $p = 0.05$. In each experiment, the number of simulated On-Off intervals was at least 10^8, or equivalently, the length of the simulated aggregated process was $\approx 2 \times 10^9$. This was necessary to ensure the desired precision of the simulation outcomes. Unfortunately, due to the space limitations we present only one simulation example. For more examples please check [21].

Experiment 1 Choose $\alpha = 3, r = 2, c = 3$. This gives $\mathbb{E}\tau^{on} = 1.202$, and $\mathbb{E}a_t^i = 0.113$. Then, for $N = 20, 25$, sources, the approximations are given by $\beta/x^3, \beta = 4.14, 48.04$, respectively. The desirable closeness between the simulation results and the approximations is represented in Figure 1. (It is interesting to observe that in this case the peak rate of each individual source is smaller than the capacity of the server.)

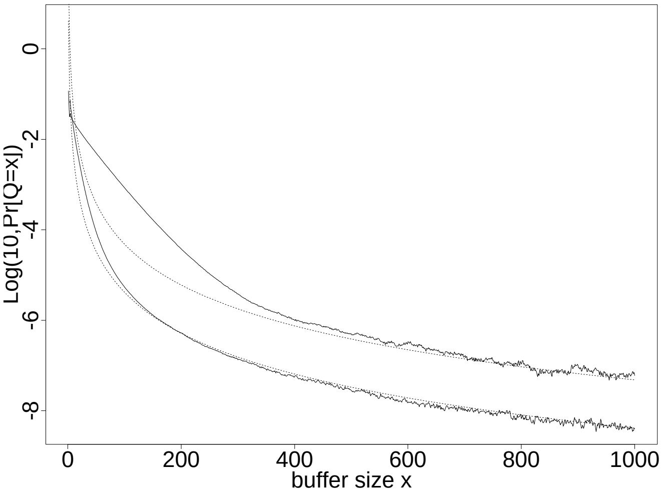

Figure 1. Illustration for Experiment 1.

Notice that in the experiment the probabilities are very small ($\approx 10^{-8}$). Hence, in order to achieve reasonable simulation accuracy, we had to choose a very large number (10^9) of simulated On-Off intervals. This means that the aggregate process was approximately 2×10^{10} samples long. The simulation of these case took *77 hours*!!! on a modern (200 MIPS) IBM workstation. *On the other hand, it is needless to say that the evaluation of (16), or (15), takes negligible time!*

5. Conclusion

In this paper, for the limiting M/G/∞ arrivals (e.g., large number of subexponential On-Off sources) with regularly varying On periods (with noninteger exponents) we obtained a precise queue asymptotics observed at the beginning of the arrival process activity periods. This showed that the asymptotic (time average) queue lower bound that was derived in [23] (under more general assumptions of intermediately varying On periods) is tight.

Based on these asymptotic results, a computationally efficient approximation was suggested for the large buffer probabilities of finitely many subexponential On-Off sources. The accuracy of this approximation was verified using extensive simulation experiments.

The results in this paper brought us closer to understanding the subexponential queueing asymptotics of multiplexed long-tailed sources. From a mathematical perspective, the elegance of the obtained results shows that long-tailed and subexponential distributions provide a proper framework for modeling and analysis of heavily dependent traffic streams. Finally, the precision and negligible computational complexity of the M/G/∞ approximation is expected to have a practical engineering impact on improving the efficacy of ATM admission controllers.

Acknowledgements

The authors are very grateful to Onno Boxma for the preprint of his paper, and to Claudia Klüppelberg for her helpful comments.

REFERENCES

1. V. Anantharam. On the sojourn time of sessions at an ATM buffer with long-range dependent input traffic. In *Proceedings of the 34th IEEE Conference on Decision and Control*, December 1995.
2. D. Anick, D. Mitra, and M. Sondhi. Stochastic theory of a data handling system with multiple sources. *Bell Syst. Techn. J.*, 61:1871–1894, 1982.
3. S. Asmussen. *Applied Probability and Queues.* Wiley, 1987.
4. S. Asmussen, L. F. Henriksen, and C. Klüppelberg. Large claims approximations for risk processes in a Markovian environment. *Stochastic Processes and their Applications*, 54:29–43, 1994.
5. F. Baccelli and P. Bremaud. *Elements of Queueing Theory: Palm-Martingale Calculus and Stochastic Recurence.* Springer Verlag, 1994.
6. N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation.* Cambridge University Press, Cambridge, 1987.
7. O. J. Boxma. Fluid queues and regular variation. *Performance Evaluation*, 27-28:699–712, 1996.
8. V. P. Chistyakov. A theorem on sums of independent positive random variables and its application to branching random processes. *Theory Probab. Appl.*, 9:640–648, 1964.
9. G. L. Choudhury and W. Whitt. Long-tail buffer-content distributions in broadband networks. preprint, 1995.
10. E. Cinlar. Superposition of point processes. In P. A. W. Lewis, editor, *Stochastic Point Processes: Statistical Analysis, Theory and Application*, pages 594–606. New York: Wiley, 1972.
11. Daren B. H. Cline. Intermediate regular and π variation. *Proc. London Math. Soc.*, 68(3):594–616, 1994.
12. J. W. Cohen. Some results on regular variation for distributions in queueing and fluctuation theory. *J. Appl. Probab.*, 10:343–353, 1973.
13. J. W. Cohen. Superimposed renewal processes and storage with gradual input. *Stochastic Process. Appl.*, 2:31–58, 1974.

14. J. W. Cohen. On the effective bandwidth in buffer design for the multi-server channels. Technical report, CWI Report BS-R9406, 1994.
15. D. R. Cox and W. L. Smith. On the superposition of renewal processes. *Biometrika*, 41:91–99, 1954.
16. N. G. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single-server queue with applications. *Mathematical Proceedings of the Cambridge Philosophical Society*, 118:363–374, 1995.
17. A. Elwalid, D. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss. Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing. *IEEE Journal on Selected Areas in Communications*, 13(6):1004–1016, August 1995.
18. P. V. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. In J. Galambos and J. Gani, editors, *Studies in Applied Probability*, volume 31A (special issue of *J. Appl. Probab.*), pages 131–156. Applied Probability Trust, Sheffield, England, 1994.
19. D. Heath, S. Resnick, and G. Samorodnitsky. Heavy tails and long range dependence in on/off processes and associated fluid models. preprint, 1996.
20. D. P. Heyman and T. V. Lakshman. Source models for VBR broadcast-video traffic. *IEEE/ACM Trans. Networking*, 4(1):40–48, 1996.
21. P. R. Jelenković and A. A. Lazar. Asymptotic results for multiplexing subexponential on-off processes CTR Technical Report CU/CTR/TR 457-96-23, Columbia University, June 1996. (www: http://www.ctr.columbia.edu/comet/publications).
22. P. R. Jelenković and A. A. Lazar. A network multiplexer with multiple time scale and subexponential arrivals. In P. Glasserman, K. Sigman, and D. D. Yao, editors, *Stochastic Networks: Stability and Rare Events*, Lecture Notes, pages 215–235. Springer-Verlag, 1996.
23. P. R. Jelenković and A. A. Lazar. Multiplexing on-off sources with subexponential on periods: Part I. In *Proc. IEEE Infocom*, Kobe, Japan, April 1997.
24. P. R. Jelenković and A. A. Lazar. Subexponential asymptotics of a Markov-modulated random walk with queueing applications. *J. Appl. Probab.*, 35(2), 1998, to appear.
25. P. R. Jelenković, A. A. Lazar, and N. Semret. Multiple time scales and subexponentiality in MPEG video streams. In *International IFIP-IEEE Conference on Broadband Communications*, April 1996.
26. P. R. Jelenković, A. A. Lazar, and N. Semret. The effect of multiple time scales and subexponentiality of MPEG video streams on queueing behavior. *IEEE J. Select. Areas Commun.*, 15(6), 1997.
27. J. Karamata. Sur un mode de croissance régulière des fonctions. *Mathematica (Cluj)*, 4:38–53, 1930.
28. C. Klüppelberg. Subexponential distributions and integrated tails. *J. Appl. Probab.*, 25:132–141, 1988.
29. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic. In *Proc. ACM Sigcomm*, pages 183–193, 1993.
30. N. Likhanov, B. Tsybakov, and N. D. Georganas. Analysis of an ATM buffer with self-similar ("fractal") input traffic. In *Proc. IEEE Infocom*, pages 985–991, Boston, Masssachusetts, April 1995.
31. R. M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Proc. Cambridge Philos. Soc.*, 58:497–520, 1962.
32. I. Norros. A storage model with self-similar input. *Queueing Syst. Theory Appl.*, 16:387–396, 1994.
33. A.G. Pakes. On the tails of waiting-time distribution. *J. Appl. Probab.*, 12:555–564, 1975.
34. M. Parulekar and A. M. Makowski. Tail probabilities for M/G/∞ input processes (I): Preliminary asymptotics. preprint, 1996.
35. M. Parulekar and A. M. Makowski. Tail probabilitites for a multiplexer with self-similar traffic. In *Proc. IEEE Infocom*, San Francisco, California, March 1996.
36. S. Resnick and G. Samorodnitsky. Performance decay in a single server exponential queueing model with long range dependence. *Operations Research*, 45(2):235, 1997.
37. M. Rubinovitch. The output of a buffered data communication system. *Stochastic Process. Appl.*, 1:375–380, 1973.
38. B. K. Ryu and S. B. Lowen. Point process approaches to the modeling and analysis of self-similar traffic - part I: Model construction. In *Proc. IEEE Infocom*, San Francisco, California, March 1996.