Self-Similar Network Traffic and Performance Evaluation, Edited by Kihong Park and Walter Willinger

Print ISBN 0-471-31974-0 Electronic ISBN 0-471-20644-X

10

ASYMPTOTIC ANALYSIS OF QUEUES WITH SUBEXPONENTIAL ARRIVAL PROCESSES

P. R. JELENKOVIĆ

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

10.1 INTRODUCTION

One of the major challenges in designing modern communication networks is providing quality of service to the individual users. An important part of this design process is understanding statistical characteristics of network traffic streams and their impact on network performance. Unlike the conventional voice traffic, modern data traffic exhibits an increased level of "burstiness" that spans over multiple time scales. It was observed that sample paths of these data sequences show evidence of self-similarity. Their autocorrelation structure is characterized by long-range dependency and the empirical distributions are easily matched with subexponential and long-tailed distributions. Early discovery of the self-similar nature of Ethernet traffic was reported in Leland et al. [42] (see also Leland et al. [43]). More recently, Crovella [22] attributed the long-range dependency of Ethernet traffic to the long-tailed file sizes that are transferred over the network. Long-range dependency of the variable bit rate video traffic was demonstrated by Beran et al. [9]. Long-tailed characteristics of the scene length distribution of MPEG video streams were explored in Heyman and Lakshman [30] and Jelenković et al. [37].

Practical importance, novelty, and the intriguing nature of these phenomena have attracted a great number of scientists to develop new traffic models and to understand the impact of these models on network performance. In this development there

have been two basic approaches: self-similar (fractal) processes and fluid renewal models with long-tailed renewal distributions. In this presentation we focus on the latter. The investigation of queueing systems with self-similar arrival processes can be found in the literature [23, 24, 44, 47, 49, 51, 54, 55].

In this chapter some recent results are presented on the subexponential asymptotic behavior of queueing systems with subexponential arrival streams. The related references will be listed throughout the chapter. First, in Section 10.2 the classes of long-tailed and subexponential distributions are defined and some of their basic properties are presented. Section 10.3 begins with a presentation of a classical result on the subexponential asymptotics of a $GI/GI/1$ queue. That is followed by a brief discussion of various extensions of this result that can be found in the literature. The remainder of Section 10.3 contains two new results on this subject. In Section 10.3.1 a derivation is given for a straightforward asymptotic approximation for the loss rate in a finite buffer $GI/GI/1$ queue. It appears surprising that the derived asymptotic formula does not depend on the queue service process. However, a simple intuitive explanation of this insensitivity effect is provided. In Section 10.3.2 a $GI/GI/1$ queue with truncated heavy-tailed arrival sequences is analyzed. Explicit asymptotic characterization of a unique behavior of the queue length distribution is given. Informally, this distribution on the log scale resembles a *stair-wave* function that has steep drops at specific buffer sizes. This has important design implications, suggesting that negligible increases of the buffer size in certain buffer regions can decrease the overflow probabilities by orders of magnitude.

Section 10.4 describes a class of fluid queues and addresses the problem of multiplexing on/off sources with heavy-tailed on periods. A complete rigorous treatment of the subexponential asymptotic behavior of a fluid queue with a single on/off arrival process is presented in Section 10.4.1. Section 10.4.2 investigates multiplexing a heavy-tailed on/off process with a process that has a lighter (exponential) tail. It is shown that this queueing system is asymptotically equivalent to the queueing system in which the process with the lighter tail is replaced by its mean value. This has implications on multiplexing bursty data and video traffic with relatively smooth voice sources. Section 10.4.3 addresses the problem of multiplexing on/off sources with heavy-tailed on periods. Understanding of this problem is fundamental for achieving high network resource utilization and providing quality of service in the bursty traffic environment. Under a specific stability condition this problem admits an elegant asymptotic solution. A brief conclusion of the presentation is given in Section 10.5.

10.2 LONG-TAILED AND SUBEXPONENTIAL DISTRIBUTIONS

This section contains necessary definitions of long-tailed and subexponential distributions. An extensive treatment of subexponential distributions (and further references) can be found in Cline [17, 18] or in the recent survey by Goldie and Klüppelberg [27].

Definition 10.2.1. A distribution function F on $[0, \infty)$ is called *long-tailed* ($F \in \mathscr{L}$) if

$$\lim_{x\to\infty} \frac{1 - F(x - y)}{1 - F(x)} = 1, \quad y \in \mathbb{R}. \tag{10.1}$$

Definition 10.2.2. A distribution function F on $[0, \infty)$ is called *subexponential* ($F \in \mathscr{S}$) if

$$\lim_{x\to\infty} \frac{1 - F^{*2}(x)}{1 - F(x)} = 2, \tag{10.2}$$

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

The class of subexponential distributions was first introduced by Chistakov [15]. The definition is motivated by the simplification of the asymptotic analysis of convolution tails. The best-known examples of distribution functions in $\mathscr{S}$ (and $\mathscr{L}$) are functions of regular variation $\mathscr{R}_{-\alpha}$ (in particular, Pareto family); $F \in \mathscr{R}_{-\alpha}$ if it is given by

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

where $l(x)\colon \mathbb{R}_+ \to \mathbb{R}_+$ is a function of slow variation, that is, $\lim_{x\to\infty} l(\delta x)/l(x) = 1, \delta > 1$. These functions were invented by Karamata [38] (the main reference book is by Bingham et al. [10]). The other examples include lognormal and some Weibull distributions (see Jelenković and Lazar [36] and Klüppelberg [40]).

A few classical results from the literature on subexponential distributions follow. The general relation between $\mathscr{S}$ and $\mathscr{L}$ is presented in Lemma 10.2.3.

Lemma 10.2.3 [7]. $\mathscr{S} \subset \mathscr{L}$.

Lemma 10.2.4. *If $F \in \mathscr{L}$ then $(1 - F(x))e^{\alpha x} \to \infty$ as $x \to \infty$, for all $\alpha > 0$.*

NOTE 10.2.5. Lemma 10.2.4 clearly shows that for long-tailed distributions Cramér-type conditions are not satisfied.

One of the most basic properties of subexponential distributions is given in the following lemma. It roughly states that the sum of n i.i.d. random variables exceeds a large value x due to one of them exceeding x.

Lemma 10.2.6. *Let $\{X_n, n \geq 1\}$ be a sequence of i.i.d. random variables with a common distribution F and let $S_n = \sum_{i=1}^{n} X_i$. If $F \in \mathscr{S}$, then*

$$\mathbb{P}[S_n > x] \sim n\mathbb{P}[X_1 > x] \quad \text{as } x \to \infty. \tag{10.3}$$

Often in renewal theory it is of interest to investigate the *integrated tail* of a distribution function. To simplify the notation, for any distribution F we denote by $\bar{F}(x) = 1 - F(x)$, $\hat{F}(x) \stackrel{\text{def}}{=} \int_x^\infty \bar{F}(t)\, dt$, and $F_1(x) \stackrel{\text{def}}{=} m^{-1}(m - \hat{F}(x))$, where $m = \hat{F}(0)$. Throughout the text $F_1(x)$ will be referred as the integrated tail distribution of $F(x)$.

Definition 10.2.7. $F \in \mathscr{S}^*$ if

$$\int_0^x \frac{\bar{F}(x-y)}{\bar{F}(x)} \bar{F}(y)\, dy \to 2m_F < \infty, \quad \text{as } x \to \infty,$$

where $m_F = \int_0^\infty yF(dy)$.

This class of distributions has the property that $F \in \mathscr{S}^* \Rightarrow F_1 \in \mathscr{S}$, and that $\mathscr{S}^* \subset \mathscr{S}$. Sufficient conditions for $F \in \mathscr{S}^*$ can be found in Klüppelberg [41], where it was explicitly shown that lognormal, Pareto, and certain Weibull distributions are in $\mathscr{S}^*$.

10.3 LINDLEY'S RECURSION AND *GI*/*GI*/1 QUEUE

Let $\{A, A_n, n \in \mathbb{N}_0\}$ and $\{C, C_n, n \in \mathbb{N}_0\}$ be two independent sequences of i.i.d. random variables (on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$). We term A_n and C_n as the arrival and service process, respectively. Then, for any initial random variable Q_0, the following Lindley's equation,

$$Q_{n+1} = (Q_n + A_{n+1} - C_{n+1})^+, \tag{10.4}$$

defines the *discrete-time queue length process* $\{Q_n, n \geq 0\}$. According to the classical result by Loynes [45] (see also Baccelli and Bremaud [8, Chap. 2]), there exists a unique stationary solution to recursion (10.4), and for all initial conditions the queue length process converges (in finite time) to this stationary process. In this chapter it is assumed that the queue is in its stationary regime, that is, $\{Q_n, n \geq 0\}$ is the stationary solution to recursion (10.4).

Recursion (10.4) also represents the waiting time process of the $GI/GI/1$ queue with C_n being interpreted as the interarrival time between the customer $n-1$ and n, A_n as the customer's n service requirement, and Q_n as the customer's n waiting time. For that reason the terms *waiting time distribution* for the $GI/GI/1$ queue and the *queue length distribution* for the discrete time queue will be used interchangeably.

Some of the first applications of long-tailed distributions in queueing theory were done by Cohen [20] and Borovkov [11] for the functions of regular variations. Cohen derived the asymptotic behavior of the waiting time distribution for the $M/GI/1$ queue. This result was extended by Pakes [48] to $GI/GI/1$ queue and the whole class of subexponential distributions. In Veraverbeke [56] the same result was rederived using a random walk technique. Let G and G_1 represent the

distribution and its integrated tail distribution for A_n, respectively, $(G_1(x) = \int_0^x \mathbb{P}[A > u]\, du/\mathbb{E}A)$.

Theorem 10.3.1 (Pakes). *If $G_1 \in \mathscr{S}$ (or $G \in \mathscr{S}^*$), and $\mathbb{E}A_n < \mathbb{E}C_n$, then*

$$\mathbb{P}[Q_n > x] \sim \frac{1}{\mathbb{E}C_n - \mathbb{E}A_n} \int_x^\infty \mathbb{P}[A_n > u]\, du \text{ as } x \to \infty.$$

There are several natural avenues for extending this result. In Willekens and Teugels [58] and Abate et al. [1] asymptotic expansion refinements to Theorem 10.3.1 were investigated. For extensions of Theorem 10.3.1 to Markov-modulated $M/G/1$ queues see Asmussen et al. [4], and to Markov-modulated $G/G/1$ queues (equivalently random walks) see Jelenković and Lazar [36]. Further extension of these results to more general arrival processes was obtained in Asmussen et al. [6]. Recently, Asmussen et al. [5] established an asymptotic relationship between the number of customers in a $GI/GI/1$ queue and their waiting time distribution.

In the rest of this section recent results are presented on a $GI/GI/1$ queue with a finite buffer and truncated heavy-tailed arrival sequences.

10.3.1 Finite Buffer $GI/GI/1$ Queue

In engineering network switches it is very common to design them as loss systems. The main performance measures for these systems are loss probabilities and loss rates. Unfortunately, there are no asymptotic results in literature that address this problem under the assumption of long-tailed arrivals. Recently, I investigated this problem [31, 33].

Here, in Theorem 10.3.2, I present the main result from my earlier work [31]. The theorem gives an explicit asymptotic characterization of the loss rate in a finite buffer queue with long-tailed arrivals. This result, in combination with results from Jelenković and Lazar [35], yields a straightforward asymptotic formula for the loss rate in a fluid queue with long-tailed $M/G/\infty$ arrivals (for more details see Jelenković [31, 33]). In addition, I [31, 33] derived an explicit asymptotic approximation of buffer occupancy probabilities. This approximation is uniformly accurate for buffer sizes that are away from the buffer boundaries (zero and the maximum buffer size). Furthermore, as the maximum buffer size increases, the length of the buffer around the boundaries where the approximation does not apply stays constant. This precise knowledge of the buffer probabilities allows computation of various other functionals of the finite buffer queue.

The evolution of a finite buffer queue is defined with the following recursion:

$$Q_{n+1}^B = \min((Q_n^B + A_{n+1} - C_{n+1})^+, B), \quad n \geq 0,$$

where B is the buffer size. We assume that the queueing process is in its stationary regime. The loss rate is defined as

$$\lambda_{\text{loss}}^{B} \stackrel{\text{def}}{=} \mathbb{E}(Q_n^B + A_{n+1} - C_{n+1} - B)^+.$$

Theorem 10.3.2. *Let G_1 be the integrated tail distribution of A. If $G_1 \in \mathscr{S}$ and $\mathbb{E}A < \mathbb{E}C$, then*

$$\lambda_{\text{loss}}^{B} \stackrel{\text{def}}{=} \mathbb{E}(A - B)^+(1 + o(1)) \quad \textit{as } B \to \infty.$$

HEURISTIC 10.3.4. Following the general heuristics for subexponential distributions the large buffer overflow is due to one (isolated) large arrival A_n. At the moment when this happens (say, time n) the queue length process is, because of the stability condition $\mathbb{E}A < \mathbb{E}C$, typically very small in comparison to B. Similarly, C_n is much smaller than B. Hence, the amount that is lost at the time of overflow is approximately $(Q_n^B + A_{n+1} - C_{n+1} - B)^+ \approx (A_{n+1} - B)^+$.

Accuracy of Theorem 10.3.2 was demonstrated [31, 33] with many numerical and simulation experiments. Here, an example is presented.

Example 10.3.5. Take $C_n \equiv 2$ and an arrival distribution $\mathbb{P}[A = 0] = \frac{1}{2}$, $\mathbb{P}[A = i] = 0.461969/i^4$, $i > 0$, $\mathbb{E}A = 0.5553$. Then, we numerically compute the loss rates $\lambda_{\text{loss}}^{B}$ for the maximum buffer sizes $B = 100i$, $i = 1, \ldots, 7$. The results are presented with circles in Fig. 10.1. Note that for $B = 700$ we needed to solve a

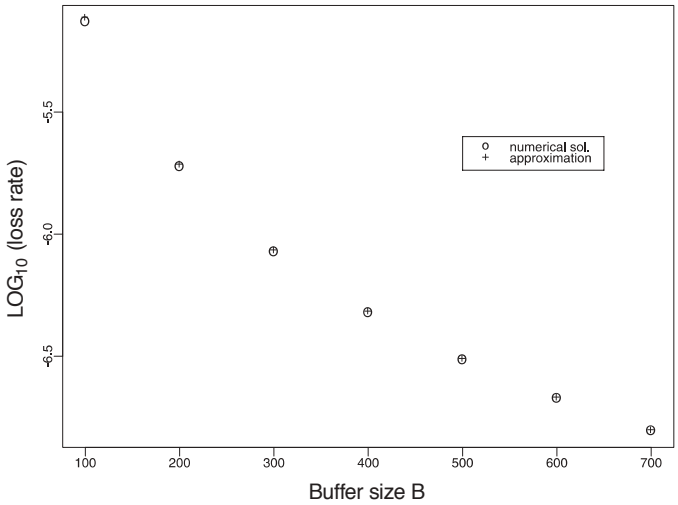

Fig. 10.1 Illustration for Example 10.3.5.

system of 700 linear equations! In contrast, Theorem 10.3.2 readily suggests an asymptotic approximation $\tilde{\lambda}^B_{\text{loss}} = 0.0767/B^2$. The approximation is presented on the same figure with "+" symbols. A precise match is apparent from the figure. In fact, relative error $|\tilde{\lambda}^B_{\text{loss}} - \lambda^B_{\text{loss}}|/\lambda^B_{\text{loss}}$ for all computed buffers was less than 4%.

10.3.2 Truncated Long-Tailed Arrival Distributions

In this section we investigate the queueing behavior when the distribution of the arrival sequence has a bounded (truncated) support [32, 34]. This arises quite frequently in practice when the arrival process distribution has a bounded support and inside that support is nicely matched with a heavy-tailed distribution (e.g., Pareto).

Our primary interest in this scenario is in its possible application to network control. More precisely, one can imagine network control procedure in which short network flows are separated from long ones. If the distribution of flows is long-tailed, this procedure will yield a truncated long-tailed distribution for the short network flows. Assume that long flows are transmitted separately using virtual circuits and short flows are multiplexed together. Intuitively, it can be expected that with short (truncated) flows one can obtain better multiplexing gains than with the original ones (before the separation). These gains are quantified in Theorem 10.3.6, which explicitly asymptotically characterizes a unique asymptotic behavior of the queue length distribution. Informally, this distribution on the log scale resembles a *stair-wave* function that has steep drops at specific buffer sizes (see Fig. 10.2). This has important design implications suggesting that negligible increases of the buffer

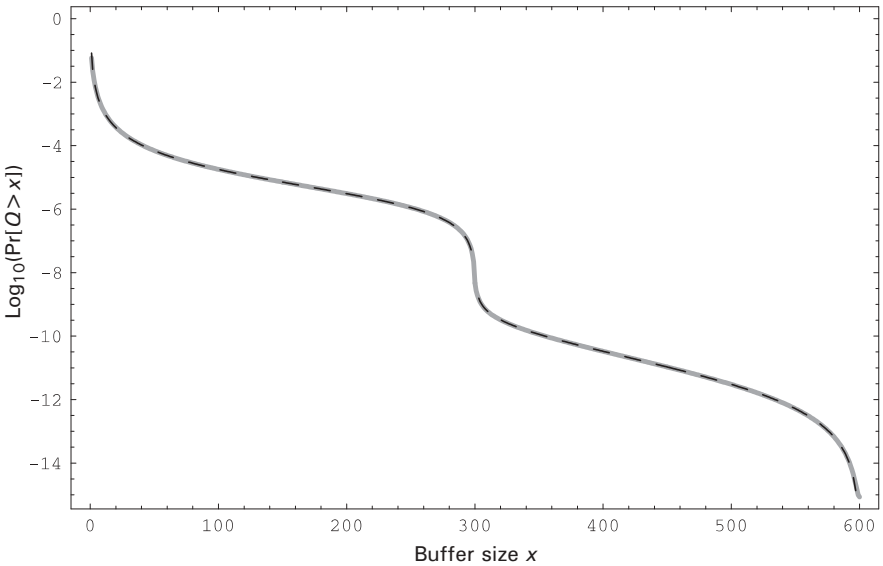

Fig. 10.2 Illustration for Example 10.3.9.

size in certain buffer regions can decrease the overflow probabilities by orders of magnitude.

Formally, for each $B > 0$ construct a sequence of truncated random variables

$$A_n^B = \min(A_{n_1} B).$$

Next, consider a single server queue with the arrival process $\{A^B, A_n^B, n \geq 0\}$, that is,

$$Q_{n+1}^B = (Q_n^B + A_{n+1}^B - C_{n+1})^+. \tag{10.5}$$

Assume that for all B, Q_n^B is in its stationary regime.

Theorem 10.3.6. *If* $\mathbb{E}(A - C) < 0$, *for all* $n > 0$, $\mathbb{P}[C > x] \leq e^{-\eta x}$, $\eta > 0$, *and* A *has a regularly varying distribution* $\mathbb{P}[A > x] = l(x)/x^\alpha$, *then*

$$\mathbb{P}[Q^B > (k + \delta)B] = \frac{h_k(\delta)}{(\mathbb{E}C - \mathbb{E}A)^{k+1}} \frac{l(B)^{k+1}}{B^{(k+1)(\alpha-1)}}(1 + o(1)) \quad \text{as } B \to \infty, \tag{10.6}$$

where $h_k(\delta)$, $0 < \delta < 1$, $k = 0, 1, 2, \ldots$, *are easily computable from*

$$h_k(\delta) \stackrel{\text{def}}{=} \int\limits_{\substack{0 < x_i \leq 1, 1 \leq i \leq k+1 \\ x_1 + \cdots + x_{k+1} \geq \delta}} x_1^{-\alpha} \cdots x_{k+1}^{-\alpha} \, dx_1 \cdots dx_{k+1}. \tag{10.7}$$

HEURISTIC 10.3.7. In order that the queue exceeds a large buffer size $b = (k + \delta)B$ it is needed that exactly $k + 1$ large arrivals (of the order B) occur at approximately the same time. Since successive arrivals are independent this event is of the order $l(B)^{k+1}/B^{(k+1)(\alpha-1)}$. The detailed proof of this result can be found in Jelenković [32].

REMARK 10.3.8. (i) This result is related to Proposition 1 in Resnick and Samorodnitsky [50], where, under conditions similar to our theorem, a rough bound for the queue length increment during an activity period of the $M/GI/\infty$ arrival process was derived. (ii) Note that $h_0(\delta)$ is explicitly given by

$$h_0(\delta) = \frac{1}{(\alpha - 1)\delta^{\alpha-1}}(1 - \delta^{\alpha-1}). \tag{10.8}$$

Now, we illustrate Theorem 10.3.6 with the following example (for more examples see Jelenković [32, 34]).

Example 10.3.9. Parameterize the distribution of A_n^B as $a_0^B = 1 - p$, $a_i^B = pd/i^{\alpha+1}$, $1 \leq i \leq B - 1$, $a_B^B = 1 - \sum_{i=0}^{B-1} a_i$, where $d = 1/\zeta(\alpha + 1)$ and $\zeta(x)$ is a Zeta function. For the choice of arrival parameters $B = 300$, $\alpha = 2.8$, and $p = 0.3$ we compute $d = 1/\zeta(\alpha + 1) = 0.273345$, $a_0^B = 0.7$, $a_i^B = 0.0820/i^{\alpha+1}$, $1 \leq i \leq B - 1$, $\rho^B = 0.34086$. For these values we numerically invert the z-transform of the queue

length distribution. These exact values of $\mathbb{P}[Q^B > x]$ are plotted with a gray line in Fig. 10.2. The values of approximation (10.6) are plotted on the same figure with dashed black lines. From the figure we can easily see that the approximation is almost identical to the exactly computed probabilities.

10.4 FLUID QUEUES AND MULTIPLEXING

Fluid queues with long-tailed characteristics have received significant attention in the recent queueing literature. The latest survey of the subject can be found in Boxma and Dumas [14]. In this section some results from Jelenković and Lazar [35] are presented.

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 the service process, respectively. Then, the evolution of the amount of fluid Q_t (also called queue length) evolves according to

$$dQ_t = (a_t - c_t)\, dt \quad \text{if } Q_t > 0, \text{ or } a_t > c_t, \tag{10.9}$$

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 Eq. (10.9) is given by

$$Q_t = \sup_{0 \leq u \leq t} \int_u^t (a_u - c_u)\, du. \tag{10.10}$$

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

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

where $W_t \stackrel{\text{def}}{=} \int_{-t}^0 (a_u - c_u)\, du,\ t \geq 0$. 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; that is,

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

Furthermore, when the difference process $x_t \stackrel{\text{def}}{=} a_t - c_t$ is driven by a stationary and ergodic point process $\{T_n,\ -\infty < n < \infty\}$, that is,

$$x_t = x_{T_n}, \quad t \in [T_n, T_{n+1}),$$

then the fluid queue process evolves as

$$Q_t = (Q_{T_{n-}} + (t - T_n)x_{T_n})^+, \quad t \in [T_n, T_{n+1}), \tag{10.11}$$

where $q^+ = \max(q, 0)$. From the recursion above, it is clear that the process Q_t is essentially the same as the $G/G/1$ workload process. Hence, by the fundamental stability theorem of Loynes there exists a unique stationary solution to Eq. (10.11). We assume that $\{Q_t,\ -\infty < t < \infty\}$ is that stationary solution.

10.4.1 Fluid Queue with a Single On/Off Process

This section presents a complete asymptotic analysis of a fluid queue with a single subexponential on/off arrival process. A general storage model in a two-state random environment was investigated by Kella and Whitt [39].

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

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

and $a_t = 0$ otherwise.

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 [8]).

$$Q_{n+1}^{\text{P}} = (Q_n^{\text{P}} + (r-c)\tau_n^{\text{on}} - c\tau_n^{\text{off}})^+, \quad n \geq 0. \tag{10.12}$$

Let F and F_1 be the distribution and the integrated tail distribution, respectively, of τ^{on}.

Theorem 10.4.1. *If $r > c$, $(r-c)\mathbb{E}\tau_{\text{on}} < c\mathbb{E}\tau_{\text{off}}$, and $F_1 \in \mathscr{S}$ (or $F \in \mathscr{S}^*$), then*

$$\mathbb{P}[Q_n^{\text{P}} > x] \sim \frac{r-c}{c\mathbb{E}\tau_{\text{off}} - (r-c)\mathbb{E}\tau_{\text{on}}} \int_{x/(r-c)}^{\infty} \mathbb{P}[\tau^{\text{on}} > u]\, du \quad \text{as } x \to \infty. \tag{10.13}$$

Proof. Define $A_n = (r-c)\tau_n^{\text{on}}$ and $C_n = c\tau_n^{\text{off}}$ and apply Theorem 10.3.1. ■

10.4.1.1 Time Averages. At this point, we will compute queue time averages based on the queue Palm probabilities computed in Theorem 10.4.1. For this we need a stationary version a_t^s of the on/off arrival process a_t. Let T_n^{on}, $-\infty < n < \infty$, be a stationary point process that represents the beginnings of the on/off periods, with a convention that $T_0^{\text{on}} < 0 \leq T_1^{\text{on}}$. Then, according to Resnick and Samorodnitsky [51], the random variable T_0^{on} can be represented as $-T_0^{\text{on}} = B(\tau_{(0)}^{\text{off}} + \tau_0^{\text{on}}) + (1-B)\tau_{(0)}^{\text{on}}$, where the random variables B, $\tau_{(0)}^{\text{on}}$, $\tau_{(0)}^{\text{off}}$ are independent of $\{\tau_n^{\text{on}}, \tau_n^{\text{off}}, n \leq -1\}$, τ_0^{off}, B is a Bernoulli random variable with $\mathbb{P}[B=0] = 1 - \mathbb{P}[B=1] = \mathbb{E}\tau^{\text{on}}/(\mathbb{E}\tau^{\text{on}} + \mathbb{E}\tau^{\text{off}})$, and $\tau_{(0)}^{\text{on}}$, $\tau_{(0)}^{\text{off}}$ are distributed as integrated tail distributions of τ^{on}, τ^{off}, respectively. Furthermore, the net increment

of the load that comes to the queue in the interval $[T_0, 0]$ is given by the following equation:

$$\int_{T_0^{\text{on}}}^{0} (a_t^s - c)\, dt = B[(r-c)\tau_0^{\text{on}} - c\tau_{(0)}^{\text{off}}] + (1-B)(r-c)\tau_{(0)}^{\text{on}}. \tag{10.14}$$

Theorem 10.4.2. *If $r > c$, $(r-c)\mathbb{E}\tau_{\text{on}} < c\mathbb{E}\tau_{\text{off}}$, and $F_1 \in \mathscr{S}$ (or $F \in \mathscr{S}^*$), then*

$$\mathbb{P}[Q_t > x] \sim \mathbb{P}[Q_n^{\text{P}} > x] + \frac{1}{\mathbb{E}\tau^{\text{off}} + \mathbb{E}\tau^{\text{on}}} \int_{x/(r-c)}^{\infty} \mathbb{P}[\tau^{\text{on}} > u]\, du \tag{10.15}$$

$$\sim K \int_{x/(r-c)}^{\infty} \mathbb{P}[\tau^{\text{on}} > u]\, du \quad \text{as } x \to \infty, \tag{10.16}$$

where

$$K = \frac{r-c}{c\mathbb{E}\tau_{\text{off}} - (r-c)\mathbb{E}\tau_{\text{on}}} + \frac{1}{\mathbb{E}\tau^{\text{off}} + \mathbb{E}\tau^{\text{on}}}. \tag{10.17}$$

REMARK 10.4.3. (i) This theorem improves on known results [16, 51] that were obtained under the assumption of τ^{on} being regularly varying. (ii) The following 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.

Most of the results in this chapter can be found elsewhere and therefore these proofs are omitted. Here, as an illustration of a subexponential proving technique, the following proof of Theorem 10.4.2 is presented. This proof is taken from Jelenković and Lazar [35].

Proof. Let $\{Q_t,\ -\infty < t < \infty\}$ be a unique stationary solution to Eq. (10.12). Then, by using Eq. (10.14), and the independence of B of Q_{T_0}, $\tau_{(0)}^{\text{off}}$, $\tau_{(0)}^{\text{on}}$, τ_0^{off}, we obtain

$$\begin{aligned}
\mathbb{P}[Q_0 > x] &= \mathbb{P}[Q_0 > x, B = 1] + \mathbb{P}[Q_0 > x, B = 0] \\
&= \mathbb{P}[Q_{T_0} + \tau_0^{\text{on}}(r-c) - c\tau_{(0)}^{\text{off}} > x, B = 1] \\
&\quad + \mathbb{P}[Q_{T_0} + (r-c)\tau_{(0)}^{\text{on}} > x, B = 0] \\
&= \frac{\mathbb{E}\tau^{\text{off}}}{\mathbb{E}\tau^{\text{on}} + \mathbb{E}\tau^{\text{off}}} \mathbb{P}[Q_{T_0} + \tau_0^{\text{on}}(r-c) - c\tau_{(0)}^{\text{off}} > x] \\
&\quad + \frac{\mathbb{E}\tau^{\text{on}}}{\mathbb{E}\tau^{\text{on}} + \mathbb{E}\tau^{\text{on}}} \mathbb{P}[Q_{T_0} + (r-c)\tau_{(0)}^{\text{on}} > x].
\end{aligned} \tag{10.18}$$

(Note that $Q_0^{\text{P}} = Q_{T_0}$). Since Q_{T_0} and $\tau_{(0)}^{\text{on}}$ are independent and subexponential and have asymptotically proportional (equivalent) tails, by applying Lemma 5(ii)(A) of

Jelenković and Lazar [35], it follows that

$$\mathbb{P}[Q_{T_0} + (r-c)\tau_{(0)}^{\text{on}} > x] \sim \mathbb{P}[Q_{T_0} > x] + \mathbb{P}[(r-c)\tau_{(0)}^{\text{on}} > x] \quad \text{as } x \to \infty. \tag{10.19}$$

The independence of τ_0^{on} and $\tau_{(0)}^{\text{off}}$, and $\tau_0^{\text{on}} \in \mathscr{L}$, by the definition of long-tailed distributions it follows that $\mathbb{P}[\tau_0^{\text{on}}(r-c) - c\tau_{(0)}^{\text{off}} > x] \sim \mathbb{P}[\tau_0^{\text{on}}(r-c) > x] = o(\mathbb{P}[Q_{T_0} > x])$ as $x \to \infty$. Subsequently, by applying Lemma 5(i)(A) of Jelenković and Lazar [35],

$$\mathbb{P}[Q_{T_0} + \tau_0^{\text{on}}(r-c) - c\tau_{(0)}^{\text{off}} > x] \sim \mathbb{P}[Q_{T_0} > x]) \quad \text{as } x \to \infty. \tag{10.20}$$

Finally, by replacing asymptotic relations (10.19) and (10.20) in Eq. (10.18), we obtain Eq. (10.15); combination of Eqs. (10.13) and (10.15) gives Eq. (10.16). This completes the proof. ■

10.4.2 Asymptotic Reduced Load Equivalence

In this section we consider multiplexing one long-tailed on/off process with exponential processes in a fluid queue. In Boxma [12, 13], a precise asymptotics of the embedded queue distribution was obtained for multiplexing on/off sources, one of which had regularly varying on periods, while the others had exponentially distributed on periods. A similar setting with intermediately varying on periods was investigated in Rolski et al. [52]. Jelenković and Lazar [35] observed that this queueing system is asymptotically interchangeable with a queueing system in which the on/off process is arriving alone and the exponential processes are replaced by their mean values. This result has been generalized in Agrawal et al. [2]. The title of this subsection is borrowed from the title of their paper.

In the remainder of this section, a result from Jelenković and Lazar [35] is presented. In order to state the result, the following definitions are introduced.

Definition 10.4.4. A distribution function F is *intermediate regular varying* $F \in \mathscr{IR}$ if

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

REMARK 10.4.5. For recent results on distributions of intermediate regular variation we refer the reader to Cline [19]. Some basic properties of $\mathscr{IR}$ are: $\mathscr{IR} \subset \mathscr{S}$; $\mathscr{R} \subset \mathscr{IR}$. Also, it is not very difficult to see that $\mathscr{IR} \subset \mathscr{S}^*S$. Therefore, all of the results obtained in this chapter apply for $\mathscr{IR}$. In addition, directly from the definition it can be shown that $F \in \mathscr{IR}$ $\int_0^\infty \bar{F}(t)\,dt < \infty, \Rightarrow F_1 \in \mathscr{IR}$.

Under the general large deviation Gärtner–Ellis conditions (see Weiss and Shwartz [57]) on the arrival process, it can be proved that the queue length distribution is exponentially bounded. To avoid stating Gärtner–Ellis conditions, we will define an arrival process e_t to be *exponential*, if whenever this process is fed

into a constant server fluid queue, the queue length distribution is exponentially bounded.

Definition 10.4.6. We say that a stationary and ergodic arrival process e_t is *exponential* if for any server capacity $c > \mathbb{E}e_t$ there exists $K \equiv K(c)$ and $\delta \equiv \delta(c) > 0$ such that

$$\mathbb{P}\left[\sup_{t\geq 0} \int_{-t}^{0} (e_u - c)\, du > x\right] \leq Ke^{-\delta x}.$$

REMARK 10.4.7. The main examples when the conditions of this definition are satisfied (i.e., Gärtner–Ellis conditions hold) are finite state space Markov chains or processes. Also, in terms of the on/off processes the conditions will hold whenever the distribution of on periods is exponentially bounded and off periods have a finite mean.

Recall that F and F_1 represent the distribution and the integrated tail distribution, respectively, of an on period. Then, we arrive at the following result (see Jelenković and Lazar [35]).

Theorem 10.4.8. *Consider a single server queue with a capacity c, and two independent arrival streams e_t and a_t. Assume that e_t is an exponential process and a_t is an on/off process with rate r, $F \in \mathscr{IR}$, and generally distributed off periods with a finite mean. If $\mathbb{E}(e_t + a_t) < c$, $r > c' \stackrel{\text{def}}{=} c - \mathbb{E}e_t$, then the queue asymptotics of this queueing system is equal to the queue asymptotics in which only the on/off process arrives and the server capacity is replaced by c', that is, it is given by Eq. (10.16) in which c is replaced by c'.*

REMARK 10.4.9. This result is true with exactly the same proof if the assumption of e_t being exponential is replaced with $\mathbb{P}[\sup_{t\geq 0} \int_{-t}^{0}(e_u - c)\, du > x] = o(F_1(x))$, for all $c > \mathbb{E}e_t$.

HEURISTIC 10.4.10. Large buildups in this fluid queue occur due to long and isolated on periods in a_t. During these long on periods the fluctuations in the exponential arrival stream e_t average out, and therefore its contribution to the asymptotic behavior is only through its mean value.

10.4.3 Multiplexing On/Off Sources

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 analysis of this problem dates back to Rubinovitch [53] and Cohen [21]. Cohen obtained a complete Laplace transform solution to this problem.

However, inverting the Laplace transform is usually a very tedious process. Hence, computationally tractable exact and approximate solution techniques are needed. For Markovian (fluid) on/off processes a thorough investigation of this problem was done in Anick et al. [3]. Many other results for multiplexing Markovian on/off processes followed. These led to the *equivalent bandwidth theory* for Markovian (or in general exponentially bounded) arrival processes; extensive references can be found in Duffield and O'Connell [24], Elwalid et al. [25], and Glynn and Whitt [26].

The analysis of a fluid queue in which more than one long-tailed process 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 process may be truly innocuous (like an on/off process). The complex autocorrelation structure of the aggregate process obtained by multiplexing long-tailed on/off processes has been examined in Heath et al. [28]. General bounds for multiplexing long-tailed fluid processes have been derived in Choudhury et al. [16]. In Poisson scaling the limiting case of an infinite number of on/off processes converges to the so-called $M/GI/\infty$ process. Asymptotic results for a fluid queue with a heavy-tailed $M/GI/\infty$ arrival process have been obtained in Boxma [12, 13] and Jelenković and Lazar [35]. Recently, new results on this model have been derived in Heath et al. [29] and Resnick and Samorodnitsky [50]. For various bounds in this context see Nain et al. [46].

10.4.3.1 Activity Period of an $M/GI/\infty$ Process. Let T_n, $n \geq 0$, $-\infty < n < \infty$, be a stationary Poisson process with rate Λ. Define $A_t^\infty = \sum_{n=-\infty}^{\infty} r1(T_n \leq t < T_n + r_n^{\text{on}})$, $r > 0$. Note that A_t^∞ represents the number of customers in an $M/GI/\infty$ queue; for that reason A_t^∞ is usually called an $M/GI/\infty$ process. An important observation is that this process represents a Poisson limit of a large number of on/off processes. Hence, it can be used as a good approximation of an aggregate process obtained by multiplexing a large (finite) number of on/off processes.

An important parameter that in many ways determines the fluid queue performance is the length of the arrival process activity period. Let $I^{\infty,\text{on}}$ be a generic activity period of an $M/GI/\infty$ process.

Theorem 10.4.11. *The asymptotics of the distribution of $I^{\infty,\text{on}}$ and its integrated tail are related as follows:*

(i) *If $F_1 \in \mathscr{S}$, then*

$$\int_t^\infty \mathbb{P}[I^{\infty,\text{on}} > u]\, du \sim e^{\Lambda \mathbb{E}\tau^{\text{on}}} \int_t^\infty \mathbb{P}[\tau^{\text{on}} > u]\, du \quad \text{as } t \to \infty.$$

(ii) *If in addition $F \in \mathscr{S}^*$, then*

$$\mathbb{P}[I^{\infty,\text{on}} > t] \sim e^{\Lambda \mathbb{E}\tau^{\text{on}}} \mathbb{P}[\tau^{\text{on}} > t] \quad \text{as } t \to \infty.$$

REMARK 10.4.12. For the case of τ^{on} being regularly varying $\mathbb{P}[\tau^{\text{on}} > t] = l(t)/t^{\alpha}$, $1 < \alpha < 2$, this result was obtained in Boxma [12] where Karamata's Tauberian/Abelian theorems were used to asymptotically relate $I^{\infty,\text{on}}$ and τ^{on}.

HEURISTIC 10.4.13. Cohen [21] shows that the expected number $\mathbb{E}N^{\infty}$ of on periods in one activity period is $e^{\Lambda \mathbb{E}\tau^{\text{on}}}$. Hence, by using the basic heuristics that a long period $I^{\infty,\text{on}}$ occurs due to exactly one long period we can jump into the conclusion

$$\mathbb{P}[I^{\infty,\text{on}} > t] \sim \mathbb{E}N^{\infty}\mathbb{P}[\tau^{\text{on}} > t] = e^{\Lambda \mathbb{E}\tau^{\text{on}}}\mathbb{P}[\tau^{\text{on}} > t].$$

However, it requires much more to rigorously prove this theorem (see Jelenković and Lazar [35]).

10.4.3.2 Queue Increment During an 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, that is, $B_n = \int_{t_n^b}^{t_n^e} A_t^{\infty}\, dt$, where t_n^b and t_n^e represent the beginning and end of the nth activity period, respectively. Furthermore, define $D_{c,n} \stackrel{\text{def}}{=} B_n - c l_n^{\text{on}}$, $0 < c \leq r$; note that $D_n \equiv D_{c,n}$ is a nonnegative 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 during the nth activity period. In order to derive the queueing asymptotics, we first have to understand the asymptotic behavior of D_n. A proof of the following result can be found in Jelenković and Lazar [35].

Theorem 10.4.14. *Consider an $M/GI/\infty$ arrival process with on periods being regularly varying $\mathbb{P}[\tau^{\text{on}} > x] = l(x)/x^{\alpha}$, $\alpha > 1$, where α is noninteger. If $0 < c \leq r$, then*

$$\mathbb{P}[D_n > x] \sim e^{\Lambda \mathbb{E}\tau^{\text{on}}}\mathbb{P}\left[\tau^{\text{on}} > \frac{x}{r + r\Lambda \mathbb{E}\tau^{\text{on}} - c}\right] \quad \text{as } x \to \infty. \tag{10.21}$$

REMARK 10.4.15. Recently in Resnick and Samorodnitsky [50] it was shown that this result holds under a more general condition of τ^{on} being intermediately regularly varying and $0 < c \leq r + r\Lambda \mathbb{E}\tau^{\text{on}}$.

10.4.3.3 Queueing Asymptotics. Let $Q_n^{\text{P},\infty}$ be the queue size observed at the beginning of the nth activity period of the $M/GI/\infty$ arrival process.

Theorem 10.4.16. *Let $\rho = \mathbb{E}A_t^{\infty} = \Lambda r \mathbb{E}\tau^{\text{on}} < c$. If $c \leq r$, and τ^{on} is regularly varying with noninteger exponent $\alpha > 1$, then*

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

Proof. Denote with $A_n = D_{c,n}$, $C_n = cl_n^{\text{off}}$, use $\mathbb{E}(C_n - A_n) = e^{\Lambda \mathbb{E}\tau^{\text{on}}}(c-\rho)/\Lambda$ and apply Theorem 10.3.1. ■

In the next theorem, under more general assumptions, we obtain a tight lower bound for the fluid queue asymptotics with $M/GI/\infty$ arrivals (see Jelenković and Lazar [35]). For this fluid queue we denote its queue content process as Q_t^{∞}. It was conjectured [35] that the following bound represents actually the exact asymptotics.

Theorem 10.4.17. *Let* $\rho \stackrel{\text{def}}{=} \mathbb{E}A_t^{\infty,s} = \Lambda r \mathbb{E}\tau^{\text{on}} < c$. *If* $r + \rho > c$, *and* $\tau^{\text{on}} \in \mathscr{IR}$, *then*

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

10.4.3.4 $M/G/\infty$ Approximation: Simulation Results. Based on Theorems 10.4.16 and 10.4.17, it is suggested that the queueing probabilities obtained by multiplexing N long-tailed on/off processes a_t^i, $1 \leq i \leq N$, are approximated as

$$\mathbb{P}[Q_t^N > x] \approx \frac{\Lambda_N r}{c_N} \int_{x/(r-c_N)}^{\infty} \mathbb{P}[\tau^{\text{on}} > u]\, du, \tag{10.22}$$

where $c_N \stackrel{\text{def}}{=} c - N\mathbb{E}a_t^i$, and $\Lambda_N \stackrel{\text{def}}{=} N\mathbb{E}a_t^i/(r\mathbb{E}\tau^{\text{on}})$. This approximation is termed an $M/G/\infty$ 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 $\mathbb{P}[\tau^{\text{off}} = t] = p(1-p)^{t-1}$, $t = 1, 2, 3, \ldots$. For on periods we consider the Pareto family $\mathbb{P}[\tau^{\text{on}} \geq t] = 1/t^{\alpha}$, $t = 1, 2, \ldots, \alpha > 0$. Here, for the discrete Pareto case we use

$$\mathbb{P}[Q_t^N = x] \approx \frac{\Lambda_N r}{c_N}(r - c_N)^{\alpha-1} x^{-\alpha}, \tag{10.23}$$

where c_N, and Λ_N are as defined earlier.

The efficacy of the approximation (10.23) is illustrated in the following simulation experiment (for additional experiments see Jelenković and Lazar [35]).

Example 10.4.18. Choose $p = 0.05$, $\alpha = 3$, $r = 2$, $c = 3$. This gives $\mathbb{E}\tau^{\text{on}} = 1.202$, and $\mathbb{E}a_t^i = 0.113$. Then, for $N = 20, 25$ processes, the approximations are given by β/x^3, $\beta = 4.14, 48.04$, respectively. The desirable closeness between the simulation results and the approximations is represented in Fig. 10.3. It is interesting to observe that in this case the peak rate of each individual process is smaller than the capacity of the server.

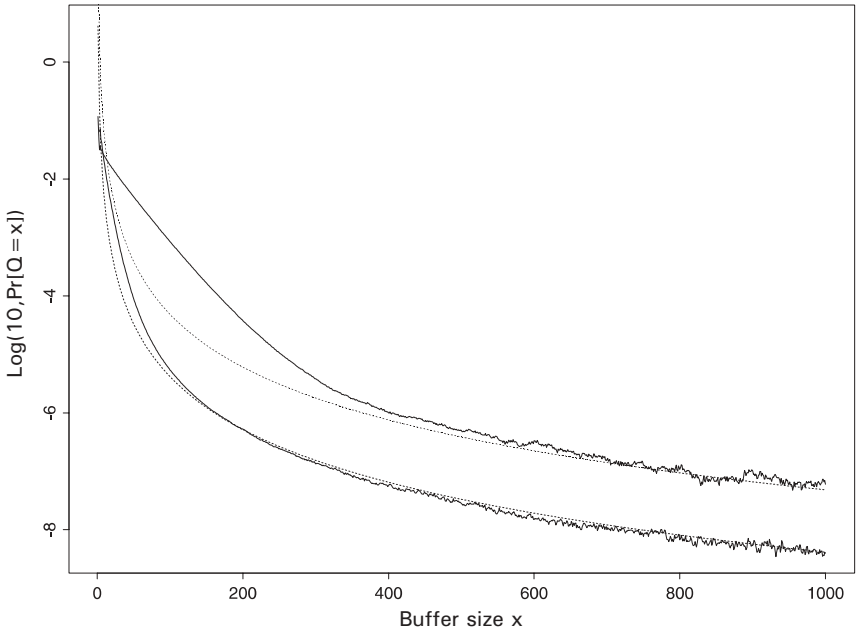

Fig. 10.3 Illustration for Example 10.4.18.

Note that in this experiment, for the case of $N = 20$ processes, 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 this case took *77 hours* on a modern (200 MIPS) IBM workstation. On the other hand, it is needless to say that the evaluation of Eq. (10.23), or Eq. (10.22), only takes a *negligible amount of time*!

10.5 CONCLUSION

In this chapter a variety of asymptotic results for queues with subexponential characteristics were presented. All of these results are *explicit*, *insightful*, and, as demonstrated with numerical examples, *accurate*. Due to these desirable characteristics, these results could be of practical use in designing future communication networks that will be able to carry efficiently and reliably bursty multimedia traffic.

ACKNOWLEDGMENTS

I am very grateful to all of my colleages who have sent me preprints of their papers.

REFERENCES

1. J. Abate, G. L. Choudhury, and W. Whitt. Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Syst.*, **16**(3/4):311–338, 1994.
2. R. Agrawal, A. M. Makowski, and P. Nain. On a reduced load equivalence for fluid queues under subexponentiality. *Queueing Syst.*, 1999, to appear.
3. D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data handling system with multiple sources. *Bell Syst. Tech. J*, **61**:1871–1894, 1982.
4. S. Asmussen, L. F. Henriksen, and C. Klüppelberg. Large claims approximations for risk processes in a Markovian environment. *Stoch. Processes Applic.*, **54**:29–43, 1994.
5. S. Asmussen, C. Klüppelberg, and K. Sigman. Sampling at subexponential times, with queueing applications. Preprint, 1998.
6. S. Asmussen, H. Schmidli, and V. Schmidt. Tail probabilities for non-standard risk and queueing processes with subexponential jumps. *Adv. Appl. Probab.*, **32**(2), 1999.
7. K. B. Athreya and P. E. Ney, *Branching Processes*. Springer-Verlag, Berlin, 1972.
8. F. Baccelli and P. Brémaud. *Elements of Queueing Theory: Palm–Martingale Calculus and Stochastic Recurrence*. Springer-Verlag, Berlin, 1994.
9. J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger. Long-range dependence in variable bit-rate video traffic. *IEEE Trans. Commun.*, **43**:1566–1579, 1995.
10. N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*. Cambridge University Press, Cambridge, 1987.
11. A. A. Borovkov. *Stochastic Processes in Queueing Theory.* Springer-Verlag, Berlin, 1976.
12. O. J. Boxma. Fluid queues and regular variation. *Perf. Eval.*, **27,28**:699–712, 1996.
13. O. J. Boxma. Regular variation in a multi-source fluid queue. In *ITC 15*, pp. 391–402, Washington, DC, June 1997.
14. O. J. Boxma and V. Dumas. Fluid queues with long-tailed activity period distributions. Technical Report PNA-R9705, CWI, Amsterdam, April 1997.
15. V. P. Chistakov. A theorem on sums of independent positive random variables and its application to branching random processes. *Theor. Probab. Appl.*, **9**:640–648, 1964.
16. G. L. Choudhury and W. Whitt. Long-tail buffer-content distributions in broadband networks. *Perf. Eval.*, **30**:177–190, 1997.
17. D. B. H. Cline. Convolution tails, product tails and domains of attraction. *Probab. Theory Relat. Fields*, **72**(1):529–557, 1986.
18. D. B. H. Cline. Convolution of distributions with exponential and subexponential tails. *J. Austral. Math. Soc. Ser. A*, **43**:347–365, 1987.
19. D. B. H. Cline. Intermediate regular and π variation. *Proc. London Math. Soc.*, **68**(3):594–616, 1994.
20. J. W. Cohen. Some results on regular variation for distributions in queueing and fluctuation theory. *J. Appl. Probab.*, **10**:343–353, 1973.
21. J. W. Cohen. Superimposed renewal processes and storage with gradual input. *Stoch. Processes Applic.*, **2**:31–58, 1974.
22. M. Crovella. The relationship between heavy-tailed file sizes and self-similar network traffic. In *9th INFORMS Applied Probability Conferences*, Cambridge, MA, June 1997.
23. K. Debicki, Z. Michna, and T. Rolski. On the supremum for Gaussian processes over infinite horizon. Preprint, May 1997.

24. N. G. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single-server queue with applications. *Math. Proc. Cambridge Philos. Soc.*, **118**:363–374, 1995.
25. 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 J. Select. Areas Commun.*, **13**(6):1004–1016, August 1995.
26. P. V. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. In J. Galambos and J. Gani, eds., *Studies in Applied Probability*, Vol. 31A (special issue of *J. Appl. Probab.*), pp. 131–156. Applied Probability Trust, Sheffield, England, 1994.
27. C. M. Goldie and C. Klüppelberg. Subexponential distributions. In M. S. Taqqu, R. Adler, and R. Feldman, eds., *A Practical Guide to Heavy Tails: Statistical Techniques for Analysing Heavy Tailed Distributions*. Birkhäuser, Basel, 1997.
28. D. Heath, S. Resnick, and G. Samorodnitsky. Heavy tails and long range dependence in on/off processes and associated fluid models. Preprint.
29. D. Heath, S. Resnick, and G. Samorodnitsky. How system performance is affected by the interplay of averages in a fluid queue with long range dependence induced by heavy tails. Preprint.
30. D. P. Heyman and T. V. Lakshman. Source models for VBR broadcast-video traffic. *IEEE J. Select. Areas Commun.*, **4**:40–48, February 1996.
31. P. R. Jelenković. Subexponential loss rates in a $GI/GI/1$ queue with applications. *Queueing Syst.*, 1999, to appear.
32. P. R. Jelenković. $GI/GI/1$ queue with truncated long-tailed service times. Submitted for publication, 1999.
33. P. R. Jelenković. Long-tailed loss rates in a single server queue. In *Proc. IEEE INFOCOM'98*, pp. 1462–1469, San Francisco, April 1998.
34. P. R. Jelenković. Network multiplexer with truncated long-tailed arrival streams. In *Proc. IEEE INFOCOM'99*, pp. 625–640, New York, NY, March 1999.
35. P. R. Jelenković and A. A. Lazar. Asymptotic results for multiplexing subexponential on–off processes. *Adv. Appl. Probab.*, **31**(2), 1999.
36. P. R. Jelenković and A. A. Lazar. Subexponential asymptotics of a Markov-modulated random walk with queueing applications. *J. Appl. Probab.*, **35**(2):325–347, June 1998.
37. P. R. Jelenković, A. A. Lazar, and N. Semret. The effect of multiple time scales and subexponentially of MPEG video streams on queueing behavior. *IEEE J. Select. Areas Commun.*, **15**(6):1052–1071, August 1997.
38. J. Karamata. Sur un mode de croissance réguilière des fonctions. *Mathematica (Cluj)*, **4**:38–53, 1930.
39. O. Kella and W. Whitt. A storage model with a two-state random environment. *Oper. Res.*, **40**:257–262, 1992.
40. C. Klüppelberg. Subexponential distributions and integrated tails. *J. Appl. Probab.*, **25**:132–141, 1988.
41. C. Klüppelberg. Subexponential distributions and characterizations of related classes. *Probab. Theory Relat. Fields*, **82**:259, 1989.
42. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic. In *Proc. ACM SIGCOMM'93*, pp. 183–193, 1993.

43. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Trans. Networking*, **2**:1–15, 1994.

44. N. Likhanov, B. Tsybakov, and N. D. Georganas. Analysis of an ATM buffer with self-similar ("fractal") input traffic. In *Proc. IEEE INFOCOM'95*, pp. 985–991, Boston, April 1995.

45. R. M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Proc. Cambridge Philos. Soc.*, **58**:497–520, 1962.

46. P. Nain, Z. Liu, D. Towsley, and Z.-L. Zhang. Asymptotic behavior of a multiplexer fed by a long-range dependent process. *J. Appl. Probab.*, **36**(1).

47. I. Norros. A storage model with self-similar input. *Queueing Syst.*, **16**:387–396, 1994.

48. A. G. Pakes. On the tails of waiting-time distribution. *J. Appl. Probab.*, **12**:555–564, 1975.

49. M. Parulekar and A. M. Makowski. Tail probabilities for a multiplexer with self-similar traffic. In *INFOCOM'96*, pp. 1452–1459, San Francisco, March 1996.

50. S. Resnick and G. Samorodnitsky. Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues. Preprint, 1997.

51. S. Resnick and G. Samorodnitsky. Performance decay in a single server queueing model with long range dependence. Preprint, 1996.

52. T. Rolski, S. Schlegel, and V. Schmidt. Asymptotics of Palm-stationary buffer content distribution in fluid flow queues. *Adv. Appl. Probab.*, **31**(1):235, 1999

53. M. Rubinovitch. The output of a buffered data communication system. *Stoch. Processes Applic.*, **1**:375–380, 1973.

54. B. K. Ryu and S. B. Lowen. Point process approaches to the modeling and analysis of self-similar traffic—part I: model construction. In *INFOCOM'96*, pp. 1468–1475, San Francisco, March 1996.

55. K. P. Tsoukatos and A. M. Makowski. Heavy traffic analysis for a multiplexer driven by $M/GI/\infty$ input processes. In *ITC 15*, pp. 497–506, Washington, DC, June 1997.

56. N. Veraverbeke. Asymptotic behavior of Wiener–Hopf factors of a random walk. *Stoch. Processes Applic.*, **5**:27–37, 1977.

57. A. Weiss and A. Shwartz. *Large Deviations for Performance Analysis: Queues*, *Communications*, *and Computing*. Chapman and Hall, New York, 1995.

58. E. Willekens and J. L. Teugels. Asymptotic expansion for waiting time probabilities in an $M/G/1$ queue with long-tailed service time. *Queueing Syst.*, **10**:295–312, 1992.