Bernoulli **7**(2), 2001, 191–210

Steady-state distribution of the buffer content for M/G/∞ input fluid queues

SIDNEY RESNICK* and GENNADY SAMORODNITSKY**

[1]*School of Operations Research and Industrial Engineering, Cornell University, Ithaca NY 14853, USA. E-mail* **sid@orie.cornell.edu; ***gennady@orie.cornell.edu*

We consider a fluid queue with on periods initiated by a Poisson process and having a long-tailed distribution. This queue has long-range dependence, and we compute the asymptotic behaviour of the steady-state distribution of the buffer content. The tail of this distribution is much heavier than the tail of the buffer content distribution of a queue which does not possess long-range dependence and which has light-tailed on periods and the same traffic intensity.

Keywords: fluid queue; heavy tails; large deviations; long-range dependence; $M/G/\infty$ queue; performance of a queue; random walk; steady-state distribution

1. Introduction and preliminaries

We consider a model of a network server (multiplexer) defined as follows. Users initiate sessions according to a Poisson process with rate $\lambda > 0$. Each session lasts a random length of time with distribution F that has a finite mean μ. The lengths of different sessions are independent of each other and of the Poisson arrival process. A session generates work or traffic or fluid at unit rate, commonly measured in some units of network traffic, e.g. packets; the work that cannot be processed immediately is collected in an infinite buffer. The server is capable of processing $r > 0$ units of traffic per unit time. Denote by $X(t)$ the buffer content at time $t \geqslant 0$. The dynamics of the buffer content process $\{X(t),\ t \geqslant 0\}$ can be expressed through its connection with the process $\{N(t),\ t \geqslant 0\}$, where $N(t)$ is the number of sessions running at time t, as

$$\mathrm{d}X(t) = N(t)\mathrm{d}t - r1(X(t) > 0)\mathrm{d}t. \tag{1.1}$$

Note that the process $\{N(t),\ t \geqslant 0\}$ in (1.1) can be viewed as describing the number of customers in the system in a $M/G/\infty$ queue where the session lengths describe the service times. We refer to $N(t)$ as the number of open sessions or the number of active connections at time t.

The above model arises as a limit of models that superimpose a finite number of independent on/off sources (see Jelenković and Lazar 1999), and it is attractive both because the pace of technological progress makes it desirable to use models that do not impose an *a priori* upper limit on the number of sources that are trying to transmit over a communication network, and also because this model is, in certain respects, more tractable than the models with a finite number of sources. We refer the reader to Boxma and Dumas

1350–7265

(1998) for a survey of literature and results for models with both finite and infinite numbers of sources. See also Vamvakos and Anantharam (1998) for a related discrete-time model.

Assume that the session length distribution has a *regularly varying tail*. That is,

$$1 - F(x) = x^{-\alpha} L(x), \qquad x \to \infty, \tag{1.2}$$

where L is a slowly varying function, and $\alpha > 1$. We write $1 - F \in \mathrm{RV}(-\alpha)$ (at infinity). This assumption is a common way to model *heavy-tailed* session lengths, and several empirical studies have confirmed the realism of the heavy-tailed assumption; see Paxson and Floyd (1994), Cunha *et al*. (1995), Crovella and Bestavros (1996) or Mikosch and Samorodnitsky (2000, Section 3). The assumption $\alpha > 1$ assures a finite mean session length and hence makes it possible for the system to be stable if the service rate r is high enough. Recent studies have found empirical evidence of α-values less than 1 for the exponent of regular variation (see, for example Arlitt and Williamson 1996) and in this case the expected session length is infinite. See Resnick and Rootzén (2000) for some indication of what may happen when $\alpha < 1$. In the present paper we concentrate on the case $\alpha > 1$.

Heavy-tailed session length distributions cause both the buffer content process $\{X(t),\ t \geqslant 0\}$ and the number of running sessions process $\{N(t),\ t \geqslant 0\}$ to possess a form of long-range dependence; see Leland *et al*. (1994), Beran *et al*. (1995), Agrawal *et al*. (1999) and Heath *et al*. (1998). It is well understood that long-range dependence usually translates into deterioration of performance of the server. This is the case if one studies the steady-state distribution of the amount of work in an infinite buffer (see, for example, Choudhury and Whitt 1997; Boxma 1997; Jelenković and Lazar 1999; Liu *et al*. 1999); it is also the case if one looks at overflow of a finite buffer and correspondingly lost traffic (Zwart 2000; Heath *et al*. 1997, 1999). See also the survey in Resnick and Samorodnitsky (2000).

Under the assumption

$$r > \lambda\mu \tag{1.3}$$

which ensures that, on average, the server is capable of coping with the traffic, the buffer content process $\{X(t),\ t \geqslant 0\}$ (recall that we assume that the buffer is infinite) reaches a steady state. In the present paper we assume (1.2) and study the tail behaviour of the marginal distribution of the steady-state buffer content process (1.1). We use the large-deviation approach that proved to be fruitful in understanding the finite buffer behaviour of the queues; see, for example, Heath *et al*. (1999) or Resnick and Samorodnitsky (1999).

In related work attention has been paid to certain 'embedded' moments of time, such as the ends of sessions and the ends of busy periods of the $M/G/\infty$ queue (the times when $N(t)$ hits zero); see Cohen (1997), Boxma and Dumas (1998) and Jelenković and Lazar (1999). Most of these cited results used various Laplace transform techniques and Tauberian theorems to invert the Laplace transforms.

We use the construction (2.4) of the stationary solution of (1.1) using the reflection map (Asmussen 1987; Prabhu 1998; Whitt 1999). Once one has an integral representation of the stationary buffer content process, it is possible to use large-deviation ideas to assess the most likely way $X(t)$ can exceed a high level. Apart from taste, there are several advantages to using the large-deviation approach in comparison with Laplace transform

techniques. The large-deviation approach is probabilistic and provides insights in the behaviour of the system. Additionally, since no inversion of transforms is used, difficulties related to integer or even values of the exponent of regular variation disappear. Furthermore, starting with a representation like (2.1) one can apply the large-deviation approach to try to compute *joint probability tails* of the buffer content measurement at several different points in time and thus understand the extreme values of the content process over intervals. We leave this, however, to future research. In the present paper we deal with the probability tail of the marginal distribution of the stationary buffer content process, i.e. with evaluating the asymptotic behaviour of $P(X(0) > \gamma)$ as γ grows large.

In Section 2 we discuss somewhat informally why the large-deviation approach can be applied in this section. Section 3 gives the formal proof of our result.

2. Probability tail of the marginal distribution of the buffer content

Our approach uses the classical representation of the steady-state buffer content process,

$$X(t) = \sup_{u \geqslant -t} \int_{-u}^{t} (N(s) - r)\mathrm{d}s = \sup_{u \leqslant t} \int_{u}^{t} (N(s) - r)\mathrm{d}s, \qquad t \geqslant 0, \tag{2.1}$$

where $\{N(t), -\infty < t < \infty\}$ is the stationary process describing the number of customers in the system in the $M/G/\infty$ queue. At any time s, $N(s)$ is a Poisson random variable with mean $\lambda\mu$ and, conditionally on $N(s) = k$, the remainders of the session lengths of the k sessions present in the sytem at time s are independent and identically distributed (i.i.d.), with the common distribution given by

$$F_*(t) = \frac{1}{\mu} \int_0^t (1 - F(u))\mathrm{d}u, \qquad t \geqslant 0. \tag{2.2}$$

Observe that if (1.2) holds, then

$$\overline{F}_*(t) \sim \frac{t^{-(\alpha-1)} L(t)}{\mu(\alpha - 1)}, \tag{2.3}$$

as a consequence of Karamata's theorem (Resnick 1987).

Note that the processes $\{X(t), t \geqslant 0\}$ and $\{N(t), t \geqslant 0\}$ related by (2.1) satisfy equation (1.1). Furthermore, since the stationary process $\{N(t), t \geqslant 0\}$ is reversible, we conclude that

$$X(0) \stackrel{\mathrm{d}}{=} \sup_{u \geqslant 0} \int_0^u (N(s) - r)\mathrm{d}s. \tag{2.4}$$

For simplicity of notation we will take $X(0)$ as *defined* by the right-hand side of (2.4). We refer the reader, once again, to Asmussen (1987), Prabhu (1998) or Whitt (1999) for more details on this construction of the stationary buffer content.

For $t \geqslant 0$ we will denote by $N_0(t)$ the number of sessions arriving in the interval $(0, t]$

and still running at time t and by $N_1(t)$ the number of the $N(0)$ initial sessions still present in the system at time t. Clearly, $N(t) = N_0(t) + N_1(t)$ for $t \geqslant 0$, and the stochastic processes $\{N_0(t)\}$ and $\{N_1(t)\}$ are independent.

Under assumption (1.3) the stochastic process

$$S_u = \int_0^u (N(s) - r)\mathrm{d}s =: A(u) - ru, \qquad u \geqslant 0, \tag{2.5}$$

behaves similarly to a negative-drift random walk, but with an important difference. The increments of a random walk 'appear instantaneously', and their effect is immediate, so that a single large increment can lift the random walk to a high level. Contrary to this, a session contributing to the process $\{N(t),\ t \geqslant 0\}$ has only gradual effect. A single ongoing session effectively reduces the service rate from r to $r-1$ and hence changes the relationship between the arrival rate and the service rate as experienced by subsequently arriving sessions. In the present study we assume that

$$\lambda\mu < r < \lambda\mu + 1. \tag{2.6}$$

Since assumption (1.3) is in force, the additional restriction provided by (2.6) is only $r < \lambda\mu + 1$. The meaning of this additional assumption is that, when a session is running, all additional sessions experience an unstable queuing system, in which the server cannot cope with the offered traffic.

The event that the supremum of the negative-drift process $\{S(u),\ u \geqslant 0\}$ in (2.5) exceeds a high level γ is unlikely, and it is exactly here the logic of large deviations applies. It says that unlikely events happen in the most likely way, and in the case of heavy tails the 'most likely way' is often that of 'the least number of causes'. The applications of this approach to a random walk type of processes go back to Embrechts and Veraverbeke (1982) and earlier. Recent applications can be found, for example, in Mikosch and Samorodnitsky (1999; 2000).

Under assumption (2.6) the 'least number of causes' mentioned above turns out to be equal to 1, and this realization drives the logic of our main result, Theorem 1. There are certain technical difficulties involved in the proof, so after stating the theorem we present an outline of it, postponing the formal proof to the next section.

Theorem 1. *Under the assumption of regular variation* (1.2) *and assumption* (2.6)*, we have*

$$P(X(0) > \gamma) \sim \frac{\lambda}{\alpha - 1} \frac{(\lambda\mu + 1 - r)^{\alpha-1}}{r - \lambda\mu} \gamma^{-(\alpha-1)} L(\gamma) \tag{2.7}$$

as $\gamma \to \infty$.

Remark 1. Observe that the result of this theorem can be rewritten as

$$P(X(0) > \gamma) \sim \frac{\lambda\mu}{r - \lambda\mu} \overline{F}_*\left(\frac{\gamma}{\lambda\mu + 1 - r}\right) \tag{2.8}$$

as $\gamma \to \infty$. This was stated under somewhat stronger conditions in Theorem 13 in Jelenković and Lazar (1999), but the argument relied on a fact conjectured, but not proved, by the

authors. The proof of the lower bound, however, in Jelenković and Lazar (1999) used neither the unproved fact nor the stronger assumption and, hence, should be attributed to them. Furthermore, Theorem 9 in Jelenković and Lazar (1999) establishes a similar result for a queue with a single on/off input source.

Related work will be found in Likhanov and Mazumdar (1999). Note that the tail of the distribution of $X(0)$ is heavier than the tail of F.

We now proceed with the promised outline of the argument. Under assumption (2.6) one, and only one, 'cause' will force the stochastic process $\{S_u, u \geqslant 0\}$ in (2.5) to reach a high positive level γ. We expect that either one of the $N(0)$ sessions remaining in the system at time 0 is long enough to cause the process $\{S_u, u \geqslant 0\}$ to reach level γ, or else the initial sessions contribute practically nothing, and it is the newly arriving sessions that are counted in $\{N_0(t), t \geqslant 0\}$ that cause the sytem to get to level γ. In addition, if it is the initial sessions that make the process $\{S_u, u \geqslant 0\}$ reach level γ, the crossing is due to exactly one extraordinarily long remaining length. Exactly how long does this remaining length have to be? While one very long session is running, the process $\{S_u, u \geqslant 0\}$ experiences a temporary positive drift of $\lambda\mu + 1 - r$, and so during this time the process $\{S_u, u \geqslant 0\}$ grows almost linearly, at that rate. That is, the only very long remaining session has to have remaining lifetime of at least $\gamma/(\lambda\mu + 1 - r)$ and, therefore,

$$P(X(0) > \gamma) = P(\sup_{u \geqslant 0} S_u > \gamma) \tag{2.9}$$

$$\approx P\left(\text{one of the } N(0) \text{ initial sessions has remaining lifetime} > \frac{\gamma}{\lambda\mu + 1 - r}\right)$$

$$+ P\left(\sup_{u \geqslant 0} \int_0^u (N_0(s) - r)\mathrm{d}s > \gamma\right).$$

However,

$$P\left(\text{one of the } N(0) \text{ initial sessions has remaining lifetime} > \frac{\gamma}{\lambda\mu + 1 - r}\right)$$

$$= \sum_{j=1}^{\infty} P(N(0) = j) P\left(\text{one of the } j \text{ initial sessions has remaining lifetime} > \frac{\gamma}{\lambda\mu + 1 - r}\right)$$

$$\approx \sum_{j=1}^{\infty} P(N(0) = j) j P\left(\text{a generic session has remaining lifetime} > \frac{\gamma}{\lambda\mu + 1 - r}\right)$$

since, once again, one, and only one, initial session can have an extraordinarily long remaining lifetime. Taking into account the common distribution (2.2) of the remaining lifetimes of the initial sessions, we immediately see that

$$P\left(\text{one of the } N(0) \text{ initial sessions has remaining lifetime} > \frac{\gamma}{\lambda\mu + 1 - r}\right)$$

$$\approx \left(1 - F_*\left(\frac{\gamma}{\lambda\mu + 1 - r}\right)\right) \mathrm{E}(N(0)) \tag{2.10}$$

$$\sim \frac{\lambda}{\alpha - 1}(\lambda\mu + 1 - r)^{\alpha-1}\gamma^{-(\alpha-1)}L(\gamma)$$

as $\lambda \to \infty$ by Karamata's theorem; see, for example, Resnick (1987).

On the other hand, the event

$$\left\{\sup_{u\geqslant 0}\int_0^u (N_0(s) - r)\mathrm{d}s > \gamma\right\},$$

if it occurs at all, has to be caused by a single session, during which the system drifts upwards to the level γ. Let W be the length of this session and T be the time when this session is initiated. Clearly, W has to be big, and so T has to be big as well since the probability of an elephantine session occurring in a short time interval is negligible. By the time T that the long session is initiated, the process is already at the (negative) level $-(r - \lambda\mu)T$. Since during the life W of the long session the system experiences a temporary positive drift of $\lambda\mu + 1 - r$, the length W of this session has to be sufficient for the system to gain $\gamma + (r - \lambda\mu)T$ units of work. Summarizing,

$$P\left(\sup_{u\geqslant 0}\int_0^u (N_0(s) - r)\mathrm{d}s > \gamma\right)$$

$$\approx P(\text{there is a session of length } W \text{ arriving at time } T \text{ and such that}$$

$$W(\lambda\mu + 1 - r) > \gamma + (r - \lambda\mu)T).$$

Since the pairs $\{(T, W)\}$ of the times the sessions are initiated and their lengths form a Poisson random measure M on $\mathbb{R}_+^2$ with mean measure

$$m(\mathrm{d}t, \mathrm{d}w) = \lambda\,\mathrm{d}t\,F(\mathrm{d}w), \tag{2.11}$$

we immediately see that

$$P\left(\sup_{u\geqslant 0}\int_0^u (N_0(s) - r)\mathrm{d}s > \gamma\right) \approx P(M(A) > 0),$$

where the set A is defined by

$$A = \{(t, w) \in \mathbb{R}_+^2 \colon w(\lambda\mu + 1 - r) > \gamma + (r - \lambda\mu)t\}.$$

A trivial computation of the asymptotic behaviour of $m(A)$ then shows that

$$P\left(\sup_{u\geqslant 0}\int_0^u (N_0(s) - r)\mathrm{d}s > \gamma\right) \approx 1 - \mathrm{e}^{-m(A)} \sim m(A) \tag{2.12}$$

$$\sim \frac{\lambda}{\alpha - 1}\,\frac{(\lambda\mu + 1 - r)^{\alpha}}{r - \lambda\mu}\,\gamma^{-(\alpha-1)}L(\gamma)$$

as $\gamma \to \infty$.

Now one only has to substitute (2.10) and (2.12) into (2.9) to obtain (2.7).

Remark 2. The logic we have just used to justify informally the result of Theorem 1 is equally well believable in the general subexponential case, that is, the case when the session length distribution is assumed to be subexponential and not, necessarily, regularly varying. See, for example, Embrechts *et al*. (1979) for more information on subexponential random variables. We conjecture, therefore, that the above theorem holds (in the form of (2.8)) in the general subexponential case. Our formal proof, however, does not carry over easily to such a general case, even though it can be generalized to certain subclasses of subexponential distributions.

We also note that a similar informal discussion is possible when the assumption $r < \lambda\mu + 1$ in (2.6) fails. However, even an informal discussion becomes quite involved without the above assumption, and hence we prefer not to present it here.

We turn now to the formal proof.

3. Formal proof of Theorem 1

We will prove that

$$\limsup_{\gamma\to\infty}(\gamma^{\alpha-1}L(\gamma)^{-1})P(X(0)>\gamma) \leqslant \frac{\lambda}{\alpha-1}\,\frac{(\lambda\mu+1-r)^{\alpha-1}}{r-\lambda\mu}. \tag{3.1}$$

Since the corresponding lower bound was proved in Theorem 11 of Jelenković and Lazar (1999), this will be sufficient for the statement of our theorem. Nevertheless, the same approach used to prove (3.1) can also be used to prove the lower bound. The reader is welcome to observe what modifications in the argument below are necessary to this end.

Our presentation will be clearer with the introduction of additional notation. Define the Poisson random measure $\mathscr{M}$ on $[0, \infty) \times \mathbb{R}_+$ by

$$\mathscr{M} = \sum_{i=1}^{N(0)} \varepsilon_{(0,Y_i^*)} + \sum_{i=1}^{\infty} \varepsilon_{(\Gamma_i,Y_i)}, \tag{3.2}$$

where

- $N(0)$ is independent of $\{Y_i^*\}$,
- $N(0)$ is Poisson distributed with parameter $\lambda\mu$ and $\{Y_i^*\}$ are i.i.d. with common distribution F_*,
- $\{(N(0), \{Y_i^*\})\}$ is independent of $\{(\Gamma_i, Y_i)\}$,
- $\{\Gamma_i\}$ are the points of a homogeneous Poisson process with rate λ on $(0, \infty)$ and are independent of the i.i.d. sequence $\{Y_i\}$ which has common distribution F.

The random measure $\sum_{i=1}^{N(0)}\varepsilon_{(0,Y_i^*)}$ is a Poisson process on $\{0\} \times \mathbb{R}_+$ and has mean measure $\delta_0(\mathrm{d}t) \times F_*(\mathrm{d}y)$, while $\sum_{i=1}^{\infty}\varepsilon_{(\Gamma_i,Y_i)}$ is Poisson on $(0, \infty) \times \mathbb{R}_+$ with mean measure $\lambda L \times F$, where L stands for Lebesgue measure. With this notation, we have

$$N_1(t) = \sum_{i=1}^{N(0)} \varepsilon_{(0,Y_i^*)}(\{0\} \times (t, \infty))$$

$$= \mathscr{M}(\{0\} \times (t, \infty))$$

$$N_0(t) = \mathscr{M}(\{(u, l): 0 < u < t < u + l\}),$$

and

$$N(t) = N_1(t) + N_0(t).$$

We start with preliminary separation of the effect of the initial sessions and that of subsequently arriving sessions as in the approximate equality in (2.9). We fix an $\varepsilon \in (0, 1)$ and write

$$P(X(0) > \gamma) = P\left(\sup_{u \geqslant 0} S_u > \gamma\right) \tag{3.3}$$

$$\leqslant P\left(\bigvee_{u \geqslant 0} S_u > \gamma, \bigvee_{i=1}^{N(0)} Y_i^* > \varepsilon\gamma\right) + P\left(\bigvee_{u \geqslant 0} S_u > \gamma, \bigvee_{i=1}^{N(0)} Y_i^* \leqslant \varepsilon\gamma\right)$$

$$=: P(A_{\gamma,\varepsilon}) + P(B_{\gamma,\varepsilon}).$$

We evaluate the probability $P(A_{\gamma,\varepsilon})$, which should be viewed as describing the effect of the $N(0)$ initial sessions. Let $Y_{(i)}^*$, $i = 1, \ldots, N(0)$, be the remaining lifetimes of the $N(0)$ initial sessions arranged in non-increasing order, and let

$$R = \sum_{i=2}^{N(0)} Y_{(i)}^*$$

be the total amount of work remaining in all initial sessions but the longest. Recall that Y_i^*, $i \geqslant 1$, are i.i.d. random variables with the common law (2.2) and independent of $N(0)$. The crucial observation here is that in the case of subexponential (and, in particular, regularly varying) tails two subexpontential random variables in a Poisson sample are much less likely to be large simulateneously than than just one. One implication of that is

$$P\left(\sum_{i=1}^{N(0)} Y_i^* > \gamma\right) \sim \mathrm{E}(N(0))P(Y_1^* > \gamma) \sim \lambda F_*(\gamma) \tag{3.4}$$

(see Embrechts *et al*., 1979). Write

$$P(Y_{(1)}^* > \varepsilon\gamma,\ R > \varepsilon\gamma) \leqslant P(Y_{(1)}^* > \varepsilon\gamma,\ Y_{(2)}^* > \varepsilon^2\gamma) + P(Y_{(1)}^* > \varepsilon\gamma,\ N(0) > \varepsilon^{-1}).$$

Observe that, because $\sum_{i=1}^{N(0)} \varepsilon_{Y_i^*}(\varepsilon^2\gamma, \infty)$ is a Poisson random variable with mean $\lambda\mu\overline{F}_*(\varepsilon^2\gamma)$, we have

$$P(Y^*_{(1)} > \varepsilon\gamma,\ Y^*_{(2)} > \varepsilon^2\gamma) \leqslant P\left[\sum_{i=1}^{N(0)} \varepsilon_{Y_i^*}(\varepsilon^2\gamma, \infty) \geqslant 2\right]$$

$$= \frac{1}{2}(\lambda\mu\overline{F}_*(\varepsilon^2\gamma))^2$$

$$= o(\overline{F}_*(\gamma)),$$

where we have used the regular variation of $\overline{F}_*(\gamma)$, and thus we conclude

$$P(Y^*_{(1)} > \varepsilon\gamma,\ Y^*_{(2)} > \varepsilon^2\gamma) = o(P[Y^*_1 > \gamma]).$$

Furthermore,

$$P\left[\bigvee_{i=1}^{N(0)} Y_i^* > \varepsilon\gamma,\ N(0) > \varepsilon^{-1}\right] = \mathrm{E}1_{[N(0)>\varepsilon^{-1}]}P\left[\bigvee_{i=1}^{N(0)} Y_i^* > \varepsilon\gamma | N(0)\right]$$

$$\leqslant \mathrm{E}N(0)1_{[N(0)>\varepsilon^{-1}]}\overline{F}_*(\varepsilon\gamma),$$

and thus we conclude

$$\limsup_{\gamma\to\infty} \frac{P(Y^*_{(1)} > \varepsilon\gamma,\ R > \varepsilon\gamma)}{P(Y^*_{(1)} > \gamma)} \leqslant \varepsilon^{-(\alpha-1)}\mathrm{E}(N(0)1_{[N(0)>\varepsilon^{-1}]}) \tag{3.5}$$

$$\leqslant \varepsilon^{-(\alpha-1)}\varepsilon^k \mathrm{E}N(0)^{k+1} \to 0$$

as $\varepsilon \to 0$ if we pick $k > \alpha - 1$. This motivates the decomposition

$$P(A_{\gamma,\varepsilon}) \leqslant P(A_{\gamma,\varepsilon} \cap \{R \leqslant \varepsilon\gamma\}) + P(Y^*_{(1)} > \varepsilon\gamma,\ R > \varepsilon\gamma). \tag{3.6}$$

Throughout the proof we will repeatedly be using the following simple majorization argument. Suppose that for some $T > 0$ and $\gamma > 0$ the event

$$\left\{\int_0^T (N(s) - r)\mathrm{d}s > \gamma\right\}$$

occurs. Then for any $k \geqslant 1$, and $0 = t_0 < t_1 < \ldots < t_k = T$ and $0 \leqslant n_j \leqslant \min_{t_j < s < t_{j+1}} N(s)$ for $j = 0, 1, \ldots, k-1$, the event

$$\left\{\int_0^T n(s)\mathrm{d}s + \sup_{S\leqslant u\leqslant T}\int_0^u (N(s) - n(s) - r)\mathrm{d}s > \gamma\right\} \tag{3.7}$$

occurs as well for any $0 \leqslant S \leqslant T$. Here $n(s) = n_j$ for $t_j < s < t_{j+1}$. We refer to this as argument M. Observe that this is a sample path argument that has nothing to do with probability law governing the process $\{N(t),\ t \geqslant 0\}$. In words, argument M says that bringing work in instantaneously in any number of presently running sessions can only make it easier for the system to cross a level.

We now apply argument M to the first probability in the right-hand side of (3.6) as follows. Let P_1 be a probability measure on the underlying space $(\Omega, \mathscr{F})$ under which the

process $\{N(t),\ t \geqslant 0\}$ is initiated with a single remaining session that we will call for obvious reasons Y_1^*, whose law is given by (2.2) conditional on $Y_1^* > \varepsilon\gamma$. Using argument M to bring in instantaneously all R units of work remaining in all initial sessions but the longest, we conclude that

$$P(A_{\gamma,\varepsilon} \cap \{R \leqslant \varepsilon\gamma\}) \tag{3.8}$$

$$\leqslant \sum_{n=0}^{\infty} \mathrm{e}^{-\lambda\mu} \frac{(\lambda\mu)^n}{n!}\, n\, P(Y_1^* > \varepsilon\gamma) P_1\left(\sup_{u \geqslant 0} S_u > (1-\varepsilon)\gamma\right)$$

$$\leqslant \lambda\mu\left[P_1\left(Y_1^* > \frac{\gamma(1-2\varepsilon)}{\lambda\mu+1-r}\right) + P_1\left(\sup_{u\geqslant 0} S_u > (1-\varepsilon)\gamma,\ Y_1^* \leqslant \frac{\gamma(1-2\varepsilon)}{\lambda\mu+1-r}\right)\right] P[Y_1^* > \varepsilon\gamma].$$

We claim that

$$\lim_{\gamma\to\infty} P_1\left(\sup_{u\geqslant 0} S_u > (1-\varepsilon)\gamma,\ Y_1^* \leqslant \frac{(1-2\varepsilon)}{\lambda\mu+1-r}\gamma\right) = 0. \tag{3.9}$$

Once (3.9) has been established, we may conclude by (3.5), (3.6), (3.8) and (3.9) that

$$\limsup_{\gamma\to\infty} \frac{P(A_{\gamma,\varepsilon})}{\gamma^{-(\alpha-1)}L(\gamma)} \leqslant \frac{\lambda}{\alpha-1}\left(\frac{1-2\varepsilon}{\lambda\mu+1-r}\right)^{-(\alpha-1)} + \frac{\lambda}{\alpha-1}\varepsilon^{-(\alpha-1)}\mathrm{E}(N(0)1_{[N(0)>\varepsilon^{-1}]}), \tag{3.10}$$

which is what we estimated in (2.10) (modulo letting $\varepsilon \to \infty$, which will be done later).

To check (3.9) notice that, considering all possible lengths not exceeding $(1-2\varepsilon)\gamma/(\lambda\mu+1-r)$ of the single remaining session, we can bound above the probability in (3.9) by

$$P\left(\sup_{0\leqslant u\leqslant \frac{1-2\varepsilon}{\lambda\mu+1-r}} \sup_{t\geqslant u}\left(\int_0^u (N_0(s)+1-r)\mathrm{d}s + \int_u^t (N_0(s)-r)\mathrm{d}s\right) > (1-\varepsilon)\gamma\right)$$

$$\leqslant P\left(\sup_{0\leqslant u\leqslant \frac{1-2\varepsilon}{\lambda\mu+1-r}\gamma} \sup_{t\geqslant u}\left(\int_0^u \left(N_0(s)+1-r-\frac{\lambda\mu+1-r}{1-\varepsilon}\right)\mathrm{d}s + \int_u^t (N_0(s)-r)\mathrm{d}s\right) > \frac{\varepsilon^2}{1-\varepsilon}\gamma\right)$$

$$\leqslant P\left(\sup_{t\geqslant 0}\int_0^t (N_0(s)-cs)\mathrm{d}s > \frac{\varepsilon^2}{1-\varepsilon}\gamma\right) \to 0$$

with

$$c = \min\left(r,\ r-1+\frac{\lambda\mu+1-r}{1-\varepsilon}\right)$$

as $\gamma \to \infty$ because by (2.6) we know that $c > \lambda\mu$.

We continue with the evaluation of the probability $P(B_{\gamma,\varepsilon})$ in (3.3), and it should be viewed as describing the effect of the sessions initiated after time 0, as in (2.12). Our first step is to use argument M and bring in instantaneously at time 0 all the work $Y_{(1)}^* + R$

remaining in the $N(0)$ sessions running at time 0. Since in the subexponential case a large value of a Poisson sum is due to a large value of the largest term in the sum, we have

$$\lim_{\gamma\to\infty}\frac{P(Y^*_{(1)}+R>\gamma)}{P(Y^*_{(1)}>\gamma)}=1$$

by Embrechts *et al*. (1979). Therefore,

$$P(Y^*_{(1)}+R>\gamma,\ Y^*_{(1)}\leqslant\gamma)=o(P(Y^*_{(1)}>\gamma))=o(F_*(\gamma))$$

by (3.4), and hence

$$\begin{aligned}P(B_{\gamma,\varepsilon}) &= P\left[\sup_{u\geqslant 0}S_u>\gamma,\ \bigvee_{i=1}^{N(0)}Y_i^*\leqslant\varepsilon\gamma\right] \qquad (3.11)\\ &\leqslant P\left[Y^*_{(1)}+R+\sup_{u\geqslant 0}\int_0^u(N_0(s)-r)\mathrm{d}s>\gamma,\ \bigvee_{i=1}^{N(0)}Y_i^*\leqslant\varepsilon\gamma\right]\\ &\leqslant P[Y^*_{(1)}+R>\varepsilon\gamma,\ Y^*_{(1)}\leqslant\varepsilon\gamma]+P\left[\sup_{u\geqslant 0}\int_0^u(N_0(s)-r)\mathrm{d}s>(1-\varepsilon)\gamma\right]\\ &= P\left[\sup_{u\geqslant 0}\int_0^u(N_0(s)-r)\mathrm{d}s>(1-\varepsilon)\gamma\right]+o(\gamma^{-(\alpha-1)}L(\gamma)),\end{aligned}$$

as $\gamma\to\infty$. We claim that

$$\limsup_{\gamma\to\infty}(\gamma^{\alpha-1}L(\gamma)^{-1})P\left(\sup_{u\geqslant 0}\int_0^u(N_0(s)-r)\mathrm{d}s>\gamma\right)\leqslant\frac{\lambda}{\alpha-1}\frac{(\lambda\mu+1-r)^\alpha}{r-\lambda\mu}. \qquad (3.12)$$

Assuming that (3.12) has been proved, we will have, by (3.11) and (3.12),

$$\limsup_{\gamma\to\infty}(\gamma^{\alpha-1}L(\gamma)^{-1})P(B_{\gamma,\varepsilon})\leqslant(1-\varepsilon)^{-(\alpha-1)}\frac{\lambda}{\alpha-1}\frac{(\lambda\mu+1-r)^\alpha}{r-\lambda\mu},$$

which, together with (3.3) and (3.10), shows that

$$\begin{aligned}\limsup_{\gamma\to\infty}\frac{P[X(0)>\gamma]}{\gamma^{\alpha-1}L(\gamma)^{-1}}&\leqslant(1-2\varepsilon)^{-(\alpha-1)}\frac{\lambda}{\alpha-1}(\lambda\mu+1-r)^{\alpha-1}+\frac{\lambda}{(\alpha-1)}\varepsilon^{-(\alpha-1)}\mathrm{E}(N(0)1_{[N(0)>\varepsilon^{-1}]})\\ &\quad+(1-\varepsilon)^{-(\alpha-1)}\frac{\lambda}{\alpha-1}\frac{(\lambda\mu+1-r)^\alpha}{r-\lambda\mu},\end{aligned}$$

and so our claim (3.3) follows by letting $\varepsilon\to 0$.

Therefore, all that remains is to prove (3.12). We start with several lemmas with a clear intuitive meaning. The proofs are fairly technical and hence postponed to the next section. Our first lemma says that if the event $\{\sup_{u\geqslant 0}\int_0^u(N_0(s)-r)\mathrm{d}s>\gamma\}$ does occur, then it has to occur by a time u not much bigger than γ. The intuitive reason, of course, is that the stochastic process $\{\int_0^u(N_0(s)-r)\mathrm{d}s,\ u\geqslant 0\}$ decays roughly linearly. Formally:

Lemma 1. *We have*

$$\lim_{M\to\infty} \limsup_{\gamma\to\infty} (\gamma^{\alpha-1}L(\gamma)^{-1})P\left(\sup_{u\geq M\gamma}\int_0^u (N_0(s)-r)\mathrm{d}s > \gamma\right) = 0. \tag{3.13}$$

The next lemma formally establishes that for the event $\{\sup_{u\geqslant 0}\int_0^u (N_0(s)-r)\mathrm{d}s > \gamma\}$ to occur, at least one very long session has to be initiated by the time of crossing of level γ. It is in the same spirit as Lemma 2.7 in Mikosch and Samorodnitsky (2000). Recall the random measure $\mathscr{M}$ defined in (3.2).

Lemma 2. *For every $M>0$ there is an $\varepsilon_0>0$ such that, for all $0<\varepsilon<\varepsilon_0$*

$$\lim_{\gamma\to\infty} \frac{P\left(\sup_{0\leqslant u\leqslant M\gamma}\int_0^u (N_0(s)-r)\mathrm{d}s > \gamma,\ \mathscr{M}((0,\ M\gamma]\times(\varepsilon\gamma,\ \infty)) = 0\right)}{\gamma^{-(\alpha-1)}L(\gamma)} = 0 \tag{3.14}$$

and

$$\lim_{\gamma\to\infty} \frac{P\left(\sup_{0\leqslant u\leqslant M\gamma}(G(u)-ru) > \gamma,\ \mathscr{M}((0,\ M\gamma]\times(\varepsilon\gamma,\ \infty)) = 0\right)}{\gamma^{-(\alpha-1)}L(\gamma)} = 0. \tag{3.15}$$

For a constant $M>1$ and $\varepsilon>0$, we consider the event

$$A(\varepsilon,\ M;\ \gamma) = \left\{\sup_{0\leqslant u\leqslant M\gamma}\int_0^u (N_0(s)-r)\mathrm{d}s > \gamma,\ \mathscr{M}((0,\ M\gamma]\times(\varepsilon\gamma,\ \infty)) = 1\right\}.$$

So $A(\varepsilon,\ M;\ \gamma)$ is the event of a crossing of level γ by time $M\gamma$ while exactly one session length exceeds $\varepsilon\gamma$. Since $\mathscr{M}((0,\ M\gamma]\times(\varepsilon\gamma,\ \infty))$ has a Poisson distribution with mean $\lambda M\gamma\overline{F}(\varepsilon\gamma)$, we have

$$P[\mathscr{M}((0,\ M\gamma]\times(\varepsilon\gamma,\ \infty)) > 1] = o(\gamma^{-(\alpha-1)}L(\gamma))$$

as $\gamma\to\infty$ for all fixed M and ε, and so it follows from Lemmas 1 and 2 that (3.12) will follow once we show that (for any fixed $M\geqslant 1$)

$$\limsup_{\gamma\to\infty} (\gamma^{\alpha-1}L(\gamma)^{-1})P(A(\varepsilon,\ M;\ \gamma)) \leqslant K_\varepsilon \frac{\lambda}{\alpha-1}\frac{(\lambda\mu+1-r)^\alpha}{r-\lambda\mu} \tag{3.16}$$

for some $K_\varepsilon\to 1$ as $\varepsilon\to 0$.

Let T_* and T^* be the times when the first session of length exceeding $\varepsilon\gamma$ is initiated and is finished, respectively. Observe that T_* and T^* are well-defined random variables, and that on the event $A(\varepsilon,\ M;\ \gamma)$ they are also the times of initiation and completion of the only session of length exceeding $\varepsilon\gamma$ that is initiated in $(0,\ M\gamma]$. For a $\delta\in(0,\ 1)$ we split the event $A(\varepsilon,\ M;\ \gamma)$ into two, depending on whether or not the level $(1-\delta)\gamma$ was first

exceeded by time T^* or not. If not, we use also argument M and bring in instantaneously at time T^* all work remaining in sessions running at that time. Define

$$A_1(\varepsilon, M; \gamma) = \left\{ \left(\left[\sup_{0 \leqslant u \leqslant T^* \wedge M\gamma} \int_0^u (N_0(s) - r)\mathrm{d}s > \gamma \right] \cup \left[G(T^* \wedge M\gamma) - r(T^* \wedge M\gamma) > \gamma \right] \right) \right.$$
$$\left. \cap \left[\mathscr{M}((0, M\gamma] \times (\varepsilon\gamma, \infty)) = 1 \right] \right\},$$

and we obtain

$$P(A(\varepsilon, M; \gamma)) \leqslant P\left(\sup_{\leqslant u \leqslant M\gamma} (G(u) - ru) > \delta\gamma, \mathscr{M}((0, M\gamma] \times (\varepsilon\gamma, \infty)) = 0 \right) \tag{3.17}$$
$$+ P(A_1(\varepsilon, M/(1-\delta); (1-\delta)\gamma)).$$

Recall that $G(T)$ is the total amount of work in all the sessions that were initiated in the interval $(0, T]$. An immediate application of Lemma 2 shows that the first term on the right-hand side of (3.17) is of an order smaller than $\gamma^{-(\alpha-1)}L(\gamma)$ if ε is small enough relative to δ. Therefore, (3.16) will follow once we show that for any fixed $M > 1$,

$$\limsup_{\gamma \to \infty} (\gamma^{\alpha-1} L(\gamma)^{-1}) P(A_1(\varepsilon, M; \gamma)) \leqslant K_\varepsilon \frac{\lambda}{\alpha - 1} \frac{(\lambda\mu + 1 - r)^\alpha}{r - \lambda\mu} \tag{3.18}$$

for some $K_\varepsilon \to 1$ as $\varepsilon \to 0$ (we are abusing the notation a bit by using the same K_ε in both (3.16) and (3.18)).

We further split the event $A_1(\varepsilon, M; \gamma)$ depending on whether or not the level γ was first exceeded by time T_* or not. If not, we once again use argument M and bring in instantaneously at time T_* all work remaining in sessions running at that time. Observe that on the event $A_1(\varepsilon, M; \gamma)$ by the time $T_* \wedge M\gamma$ only sessions of length not exceeding $\varepsilon\gamma$ are initiated, and, apart from the only long session, only sessions of length not exceeding $\varepsilon\gamma$ are initiated between $T_* \wedge M\gamma$ and $M\gamma$. To make the accounting easier, we define a new process, say $\{N_0^{(1)}(t), t \geqslant 0\}$, independent of T_* and T^*, and representing the number of customers in an $M/G/\infty$ queue that starts empty at time 0, in which the customers arrive at rate $\lambda F(\varepsilon\gamma)$ and the distribution of the job lengths is given by

$$F^{(1)}(A) = \frac{F(A \cap [0, \varepsilon\gamma])}{F([0, \varepsilon\gamma])}, \qquad A \text{ Borel}.$$

We can always let this process live on some new probability space, but for simplicity of notation we will assume that it lives on the same probability space $(\Omega, \mathscr{F}, P)$. We have

$$\mathrm{P}(A_1(\varepsilon, M; \gamma)) \leqslant P(A_2(\varepsilon, M; \gamma)) + P(A_3(\varepsilon, M; \gamma)), \tag{3.19}$$

where

$$A_2(\varepsilon, M; \gamma) = \left\{ \sup_{0 \leq u \leq T_* \wedge M\gamma} \int_0^u (N_0^{(1)}(s) - r)\mathrm{d}s > \gamma \right\}$$

and

$$A_3(\varepsilon, M; \gamma) =$$

$$\left\{ T_* \leqslant M\gamma \text{ and either } G^{(1)}(T_* - rT_*) + \sup_{0 \leqslant u \leqslant T^* \wedge M\gamma - T_* \wedge M\gamma} \int_0^u (N_0^{(2)}(s) + 1 - r)\mathrm{d}s > \gamma \right.$$

$$\left. \text{or } G^{(1)}(T^* \wedge M\gamma) - r(T^* \wedge M\gamma) + (T^* \wedge M\gamma - T_*) > \gamma \right\},$$

in which we are using the obvious notation that $G^{(1)}(T)$ is the total amount of work in all the sessions of the process $\{N_0^{(1)}(t),\ t \geqslant 0\}$ that were initiated in the interval $(0,\ T]$. Furthermore, we denote for $s \geqslant 0$ by $N_0^{(2)}(s)$ the number of sessions of the process $\{N_0^{(1)},\ t \geqslant 0\}$ that arrive in the interval $(T_*,\ T_* + s]$ and are still running at time $T_* + s$. The process $\{N_0^{(2)}(t),\ t \geqslant 0\}$ is a version of the process $\{N_0^{(1)}(t),\ t \geqslant 0\}$.

Clearly, Lemma 2 implies that the first term on the right-hand of (3.19) is of an order smaller than $\gamma^{-(\alpha-1)}L(\gamma)$ if ε is small enough. Therefore, (3.18) will be proved if we show that, for all $M > 1$,

$$\limsup_{\gamma \to \infty} (\gamma^{\alpha-1} L(\gamma)^{-1}) P(A_3(\varepsilon, M; \gamma)) \leqslant K_\varepsilon \frac{\lambda}{\alpha - 1} \frac{(\lambda\mu + 1 - r)^\alpha}{r - \lambda\mu} \tag{3.20}$$

for some $K_\varepsilon \to 1$ as $\varepsilon \to 0$ (with the same abuse of notation as before).

For a $T > 0$, we denote by

$$G(T) = \sum_{0 < \Gamma_k \leqslant T} Y_k$$

the total amount of work in all the sessions that *were initiated* in the interval $(0,\ T]$. The two basic observations are the obvious bound

$$G(T) \geqslant \int_0^T N_0(s)\mathrm{d}s, \tag{3.21}$$

since the right-hand side only accounts for work accomplished by time T. We have by (3.21), for any $\delta > 0$,

$$P(A_3(\varepsilon, M; \gamma)) \leqslant 2P\left(G^{(1)}(t) - \left(\lambda\mu + \frac{\delta}{2M}\right) t > \frac{\delta}{2}\gamma \text{ for some } 0 \leqslant t \leqslant M\gamma \right) \tag{3.22}$$

$$+ P\big(T_* \leqslant M\gamma,\ ((1+\delta)\lambda\mu + 1 - r)(T^* - T_*) + G^{(1)}(T_*) - rT_* > (1-\delta)\gamma\big).$$

By Lemma 2 we have

$$\lim_{\gamma\to\infty}(\gamma^{\alpha-1}L(\gamma)^{-1})P\left(G^{(1)}(t)-\left(\lambda\mu+\frac{\delta}{2M}\right)t>\frac{\delta}{2}\gamma \text{ for some } 0\leqslant t\leqslant M\gamma\right)=0 \qquad (3.23)$$

as long as ε is small enough relative to δ, and so we only have to treat the second term on the right-hand side of (3.22), which is exactly the term corresponding to the main probability term which we estimated in (2.12).

What remains is a standard computation. Recall that T_* and T^*-T_* are independent random variables, with T_* being exponentially distributed with parameter $\lambda(1-F(\varepsilon\gamma))$, and T^*-T_* having distribution given by

$$F^{(2)}(A)=\frac{F(A\cap(\varepsilon\gamma,\infty))}{F((\varepsilon\gamma,\infty))}, \qquad A \text{ Borel.}$$

Observe that

$$\begin{aligned}P(T_*\leqslant M\gamma,\,&((1+\delta)\lambda\mu+1-r)(T^*-T_*)+G^{(1)}(T_*)-rT_*>(1-\delta)\gamma)\\ &\leqslant P(((1+\delta)\lambda\mu+1-r)(T^*-T_*)+((1+\delta)\lambda\mu-r)T_*>(1-\delta)\gamma,\,T_*\leqslant M\gamma)\\ &\quad+P(G^{(1)}(T_*)>(1+\delta)\lambda\mu T_*).\end{aligned}$$

By a straightforward application of Lemma 2 (with $r=(1+\delta/2)\lambda\mu$), we see that, for every $\delta>0$,

$$P(G^{(1)}(T_*)>(1+\delta)\lambda\mu T_*)=o(\gamma^{-(\alpha-1)}L(\gamma))$$

as $\gamma\to\infty$ as long as ε is small enough relative to δ. Observe, furthermore, that for all $\delta>0$ small enough, $(1+\delta)\lambda\mu-r<0$. We conclude that (3.20) will follow if, for every fixed $M>1$,

$$\limsup_{\gamma\to\infty}\frac{P[(\lambda\mu+1-r)(T^*-T_*)+(\lambda\mu-r)T_*>\gamma,\,T_*\leqslant M\gamma]}{\gamma^{-(\alpha-1)}L(\gamma)}\leqslant K_\varepsilon\frac{\lambda}{\alpha-1}\frac{(\lambda\mu+1-r)^\alpha}{r-\lambda\mu}, \qquad (3.24)$$

once again for some $K_\varepsilon\to 1$ as $\varepsilon\to 0$.

Denote

$$a=\frac{1}{\lambda\mu+1-r}, \qquad b=\frac{r-\lambda\mu}{\lambda(\lambda\mu+1-r)}.$$

Using regular variation and Potter's bounds (see, for example, Proposition 0.8(ii) of Resnick 1987), we see that, given $1<\alpha'<\alpha$, for all $t>0$ and γ large enough,

$$\frac{1-F(\gamma(a+bt))}{1-F(\gamma)}\leqslant C(a+bt)^{-\alpha'}$$

for some $C>0$. Therefore, regular variation, together with the dominated convergence theorem, implies that, for every $0<\varepsilon<1$,

$$P((\lambda\mu + 1 - r)(T^* - T_*) + (\lambda\mu - r)T_* > \gamma,\ T_* \leqslant M\gamma)$$

$$\leqslant \frac{1}{1 - F(\varepsilon\gamma)} \int_0^\infty \mathrm{e}^{-t} \left(1 - F\left(\frac{\gamma + (r - \lambda\mu)\dfrac{t}{\lambda(1 - F(\varepsilon\gamma))}}{\lambda\mu + 1 - r}\right)\right) \mathrm{d}t$$

$$= \gamma \int_0^\infty \exp\{-\gamma(1 - F(\varepsilon\gamma))t\}(1 - F(\gamma(a + bt)))\mathrm{d}t$$

$$\sim \gamma(1 - F(\gamma)) \int_0^\infty (a + bt)^{-\alpha}\, \mathrm{d}t$$

$$= \frac{a^{-(\alpha-1)}}{(\alpha - 1)b}\gamma(1 - F(\gamma))$$

$$= \frac{\lambda}{\alpha - 1} \frac{(\lambda\mu + 1 - r)^\alpha}{r - \lambda\mu}\gamma(1 - F(\gamma)),$$

thus proving (3.24), and hence completing the proof of the theorem.

4. Proofs of Lemmas 1 and 2

The idea behind the argument in both cases is to compare the fluid queue to an appropriate random walk with negative drift, which crosses a positive level before the fluid queue does (see, for example, (3.21)).

Proof of Lemma 1. We have by (3.21), for all $u \geqslant M\gamma$,

$$\int_0^u (N_0(s) - r)\mathrm{d}s = \int_0^{M\gamma} (N_0(s) - r)\mathrm{d}s + \int_{M\gamma}^u (N_0(s) - r)\mathrm{d}s$$

$$\leqslant G(M\gamma) - rM\gamma + \int_{M\gamma}^u (\tilde{N}0(s) - r)\mathrm{d}s,$$

where $\tilde{N}_0(s)$ is the number of sessions arriving in the interval $(M\gamma, s)$ and still running at time s. We conclude that

$$P\left(\sup_{u \geqslant M\gamma} \int_0^u (N_0(s) - r)\mathrm{d}s > \gamma\right)$$

$$\leqslant P\left(\sum_{j=1}^K Y_j - rM\gamma > -\frac{r - \lambda\mu}{2} M\gamma\right) + P\left(\sup_{u \geqslant 0} \int_0^u (N_0(s) - r)\mathrm{d}s > \gamma\left(1 + \frac{r - \lambda\mu}{2} M\right)\right),$$

where K is a Poisson random variable with mean $\lambda M\gamma$ independent of a sequence of i.i.d. random variables $Y_1, Y_2, \ldots$ with common distribution F.

Observe that by (3.21),

$$\sup_{u\geqslant 0}\int_0^u (N_0(s)-r)\mathrm{d}s \leqslant \sup_{n\geqslant 0} S_n,$$

where $S_n = Z_1 + \ldots Z_n$, $n \geqslant 0$, is a random walk with $Z_i = Y_i - r(\Gamma_i - \Gamma_{i-1})$, $i \geqslant 1$. Here Y_i is the duration of the ith session, and $\Gamma_i - \Gamma_{i-1}$ is the time gap between the instances the $(i-1)$th and the ith sessions are initiated. Therefore,

$$P\left(\sup_{u\geqslant 0}\int_0^u (N_0(s)-r)\mathrm{d}s > \gamma\left(1+\frac{r-\lambda\mu}{2}M\right)\right) \leqslant P\left(\sup_{n\geqslant 0} S_n > \gamma\left(1+\frac{r-\lambda\mu}{2}M\right)\right)$$
$$\sim \frac{1}{r/\lambda-\mu}\left(1+\frac{r-\lambda\mu}{2}M\right)^{-(\alpha-1)}\frac{1}{\alpha-1}\gamma^{-(\alpha-1)}L(\gamma)$$

as $\gamma \to \infty$, by Embrechts *et al*. (1979). We conclude that

$$\lim_{M\to\infty}\limsup_{\gamma\to\infty}(\gamma^{\alpha-1}L(\gamma)^{-1}P\left(\sup_{u\geqslant 0}\int_0^u (N_0(s)-r)\mathrm{d}s > \gamma\left(1+\frac{r-\lambda\mu}{2}M\right)\right) = 0. \tag{4.2}$$

Furthermore,

$$P\left(\sum_{j=1}^{K} Y_j - rM\gamma > -\frac{r-\lambda\mu}{2}M\gamma\right)$$
$$\leqslant P\left(K > \frac{r/\mu+\lambda}{3}M\gamma\right) + P\left(\sum_{j=1}^{[(r/\mu+\lambda)M\gamma/3]} Y_j > \frac{r+\lambda\mu}{2}M\gamma\right).$$

Since $\lambda < r/\mu$, we immediately see that, for every $M > 1$,

$$P\left(K > \frac{r/\mu+\lambda}{3}M\gamma\right) = o(\mathrm{e}^{-c\gamma})$$

as $\gamma \to \infty$ for some $c > 0$; this is a classical large-deviation bound, easily obtainable via an exponential Markov inequality. On the other hand,

$$P\left(\sum_{j=1}^{[(r/\mu+\lambda)M\gamma/3]} Y_j > \frac{r+\lambda\mu}{2}M\gamma\right) \leqslant P\left(\sum_{j=1}^{[(r/\mu+\lambda)M\gamma/3]} (Y_j-\mu) > \frac{r-\lambda\mu}{6}M\gamma\right)$$
$$\sim \frac{r/\mu+\lambda}{3}M\gamma\left(1-F\left(\frac{r-\lambda\mu}{6}M\gamma\right)\right)$$

as $\gamma \to \infty$ by, say, Nagaev (1969) or Cline and Hsing (1991). We therefore conclude that

$$\lim_{M\to\infty}\limsup_{\gamma\to\infty}(\gamma^{\alpha-1}L(\gamma)^{-1})P\left(\sum_{j=1}^{K} Y_j - rM\gamma > -\frac{r-\lambda\mu}{2}M\gamma\right) = 0, \tag{4.3}$$

and now (3.13) follows from (4.2) and (4.3). This completes the proof of Lemma 1. □

Proof of Lemma 2. By (3.21), we have that (3.14) follows from (3.15), and so it is enough to prove the latter. Clearly,

$$P\left(\sup_{0\leqslant u\leqslant M\gamma}(G(u)-ru)>\gamma,\ \mathscr{M}((0,\ M\gamma]\times(\varepsilon\gamma,\ \infty))=0\right)\leqslant P\left(\sup_{n\leqslant k(\varepsilon\gamma)}S_n>\gamma\right),$$

where $\{S_n,\ n\geqslant 0\}$ is the random walk defined in the proof of Lemma 1 above, and

$$k(\varepsilon\gamma),\ \inf\{n\geqslant 0:\ Z_n>\varepsilon\gamma\}.$$

For an $M>1$, we decompose the last probability as

$$P\left(\sup_{n\geqslant k(\varepsilon\gamma)}S_n>\gamma\right)\leqslant P\left(\sup_{0\leqslant n\leqslant\gamma/M}S_n>\gamma\right)+P\left(\sup_{\gamma/M\leqslant n\leqslant k(\varepsilon\gamma)}S_n>\gamma\right)$$

$$:=p_1(M;\ \gamma)+p_2(\varepsilon,\ M;\ \gamma).$$

However, by Lemma 2.5 of Mikosch and Samorodnitsky (2000),

$$\lim_{M\to\infty}\limsup_{\gamma\to\infty}(\gamma^{\alpha-1}L(\gamma)^{-1})p_1(M;\ \gamma)=0,$$

while by Lemma 2.7 of Mikosch and Samorodnitsky (2000), for every fixed $M>1$ and for all ε small enough,

$$\lim_{\gamma\to\infty}(\gamma^{\alpha-1}L(\gamma)^{-1}p_2(\varepsilon,\ M;\ \gamma)=0.$$

Therefore, (3.15) follows, and the proof is complete. □

Acknowledgements

The research for this paper was partially supported by NSF grants DMS-0071073 and DMI-9713549 and by NSA grant MDA904-95-H-1036 at Cornell University. We gratefully acknowledge improvements by the anonymous referee which saved more than one page of computations.

References

Agrawal, R., Makowski, A. and Nain, P. (1999) On a reduced load equivalence for fluid queues under subexponentiality. *Queueing Systems Theory Appl.*, **33**, 5–41.

Arlitt, M. and Williamson, C. (1996) Web servers workload characterization: The search for invariants (extended version). In *Proceedings of the ACM Sigmetrics International Conference on Measurement and Modeling of Computer Systems*. New York: Association for Computing Machinery.

Asmussen, S. (1987) *Applied Probability and Queues*. Chichester: Wiley.

Beran, J., Sherman, R., Willinger, W. and Taqqu, M. (1995) Long-range dependence in variable-bit-rate video traffic. *IEEE Trans. Commun.*, **43**, 1566–1579.

Boxma, O. (1997) Regular variation in a multi-source fluid queue. In V. Ramaswami and P. Wirth (eds), *Teletraffic Contributions for the Information Age*, pp. 391–402. Amterdam: North-Holland.

Boxma, O. and Dumas, V. (1998) Fluid queues with long-tailed activity period distributions. *Comput. Commun.*, **21**, 1509–1529.

Choudhury, G.L. and Whitt, W. (1997) Long-tail buffer-content distributions in broadband networks. *Perform. Eval.*, **30**, 177–190.

Cline, D. and Hsing, T. (1991) Large deviation probabilities for sums and maxima of random variables with heavy or subexponential tails. Preprint, Texas A&M University.

Cohen J. (1997) The $M/G/1$ fluid model with heavy-tailed message length distributions. Technical Report PNA-R9714, Centrum voor Wiskunde en Informatica.

Crovella, M. and Bestavros, A. (1996) Self-similarity in World Wide Web traffic: evidence and possible causes. *Perform. Eval. Rev.*, **24**, 160–169.

Cunha, C., Bestavros, A. and Crovella, M. (1995) Characteristics of www client-based traces. Preprint BU-CS-95-010, Boston University.

Embrechts, P. Veraverbeke, N. (1982) Estimates for the probability of ruin with special emphasis on the possibility of large claims. *Insurance Math. Econom.*, **1**, 55–72.

Embrechts, P., Goldie, C. and Veraverbeke, N. (1979): Subexponentiality and infinite divisibility. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **49**, 335–347.

Heath, D., Resnick S. and Samorodnitsky, G. (1997) Patterns of buffer overflow in a class of queues with long memory in the input stream. *Ann. Appl. Probab.*, **7**, 1021–1057.

Heath, D., Resnick, S. and Samorodnitsky, G. (1998) Heavy tails and long range dependence in on/off processes and associated fluid models. *Math. Oper. Res.*, **23**, 145–165.

Heath, D., Resnick, S. and Samorodnitsky, G. (1999) How system performance is affected by the interplay of averages in a fluid queue with long range dependence induced by heavy tails. *Ann. Appl. Probab.*, **9**, 352–375.

Jelenković, P. and Lazar, A. (1999) Asymptotic results for multiplexing subexponential on–off sources. *Adv. Appl. Probab.*, **31**, 394–421.

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

Likhanov, N. and Mazumdar, R. (1999) Cell loss asymptotics for buffers fed with a large number of independent stationary sources. *J. Appl. Probab.*, **36**, 86–96.

Liu, Z., Nain, P., Towsley, D. and Zhang, Z.-L. (1999) Asymptotic behavior of a multiplexer fed by a long-range dependent process. *J. Appl. Probab.*, **36**, 105–118.

Mikosch, T. and Samorodnitsky, G. (1999) Ruin probability with claims modeled by a stationary ergodic stable process. Preprint, Cornell University.

Mikosch, T. and Samorodnitsky, G. (2000) The supremum of a negative drift random walk with dependent heavy-tailed steps. *Ann. Appl. Probab.*, **10**, 1025–1064.

Nagaev, A. (1969) Limit theorems for large deviations where Cramér's conditions are violated. *Izv. Akad. Nauk UzSSR Ser. Fiz.-Mat. Nauk*, **6**, 17–22 (in Russian).

Paxson, V. and Floyd, S. (1994) Wide area traffic: the failure of Poisson modelling. *IEEE/ACM Trans. Networking*, **3**, 226–244.

Prabhu, N. (1998) *Stochastic Storage Processes: Queues, Insurance Risk, Dams, and Data Communication*. New York: Springer-Verlag.

Resnick, S. (1987) *Extreme Values, Regular Variation and Point Processes*. New York: Springer-Verlag.

Resnick, S. and Rootzén, H. (2000) Self-similar communication models and very heavy tails. *Ann. Appl. Probab.*, **10**, 753–778.

Resnick, S. and Samorodnitsky, G. (1999) Activity periods of an infinitte server queue and performance of certain heavy tailed fluid queues. *Queueing Systems Theory Appl.*, **33**, 43–71.

Resnick, S. and Samorodnitsky, G. (2000) Fluid queues, leaky buckets, on-off processes and teletraffic modeling with high variable and correlated Inputs. In K. Park and W. Willinger (eds), *Self-similar Network Traffic and Performance Evaluation*. New York: Wiley.

Vamvakos, S. and Anantharam, V. (1998) On the departure process of a leaky bucket system with long-range dependent input traffic. *Queueing Systems Theory Appl.*, **28**, 191–214.

Whitt, W. (1999) The reflection map is Lipschitz with appropriate Skorohod M-metrics. Preprint, AT&T Labs Research, Florham Park, NJ.

Zwart, A. (2000) A fluid queue with a finite buffer and subexponential input. *Adv. Appl. Probab.*, **32**, 221–243.

Received June 1999 and revised September 2000