STEADY STATE DISTRIBUTION OF THE BUFFER CONTENT FOR M/G/∞ INPUT FLUID QUEUES

SIDNEY RESNICK AND GENNADY SAMORODNITSKY

ABSTRACT. We consider a fluid queue with *ON* periods arriving according to a Poisson process and having a long–tailed distribution. This queue has long range dependence, and we compute the asymptotic behavior 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.

1. INTRODUCTION AND PRELIMINARIES

We consider a model of a network server (multiplexer) defined as follows. Sessions arrive to the server 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 of time. Denote by $X(t)$ the buffer content at time $t \geq 0$. The dynamics of the buffer content process $\{X(t),\, t \geq 0\}$ can be expressed through its connection with the process $\{N(t),\, t \geq 0\}$, where $N(t)$ is the number of sessions running at time t, as

$$dX(t) = N(t)\, dt - r\mathbf{1}(X(t) > 0)\, dt\,. \tag{1.1}$$

Note that the process $\{N(t),\, t \geq 0\}$ 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 (1996) for a survey of literature and results for both finite and infinite number of sources models. See also Vamvakos and Anantharam (1998) for a related discrete time model.

AMS 1991 *subject classifications.* Primary 90B15; secondary 60K250.

Key words and phrases. fluid queue, $M/G/\infty$ queue, heavy tails, long range dependence, performance of a queue, steady state distribution, random walk, large deviations.

Research partially supported by NSF grants DMS-97-04982 and DMI-9713549 and by NSA grant MDA904-95-H-1036 at Cornell University.

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

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

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 tails* of session length, and a number of empirical studies confirmed the heavy tailed assumption. See Paxson and Floyd (1994), Cunha et al. (1995), Crovella and Bestavros (1996) or Section 3 in Mikosch and Samorodnitsky (1998). 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. Even though it is scary to think about the implications, certain recent studies have found empirical evidence of values less than 1 of the exponent of regular variation α (e.g. Arlitt and Williamson (1996)). See Resnick and Rootzén (1998) for some indication of what may happen when $\alpha < 1$. In the present paper we concentrate on the less dramatic case $\alpha > 1$.

Heavy tailed session length distributions cause both the buffer content process $\{X(t),\, t \geq 0\}$ and the number of running sessions process $\{N(t),\, t \geq 0\}$ to possess a form of long range dependence. See Leland et al. (1994), Beran et al. (1995), Agrawal et al. (1998), 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 work in an infinite buffer (see e.g. Choudhury and Whitt (1995), Boxma (1996), Jelenković and Lazar (1999) and Liu et al. (1997)). This is also the case if one looks at overflow of a finite buffer and correspondingly lost traffic (see e.g. Zwart (1998), Heath et al. (1997b), Heath et al. (1997a)). See also a survey in Resnick and Samorodnitsky (1998).

Under the assumption

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

that ensures that, on average, the server is capable of coping with the traffic, the buffer content process $\{X(t),\, t \geq 0\}$ (recall that we assume that the buffer is infinite) reaches a steady state (constructed below). In the present paper we assume (1.2) and study the tail behavior 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 behavior of the queues; see e.g. Heath et al. (1997b) or Resnick and Samorodnitsky (1997).

Remark 1. In related work attention was paid to certain "embedded" moments of time, like the ends of sessions, the ends of busy periods of the $M/G/\infty$ queue (the times when $N(t)$ hits zero), etc. See Cohen (1997), Boxma and Dumas (1996) and Jelenković and Lazar (1999). Most of the results obtained here used various Laplace transform techniques and Tauberian theorems to invert the Laplace transforms.

In Section 2 we start by reviewing the construction of the stationary solution to (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 into the behavior 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.

Importantly, starting with a representation like (2.5) one can apply the large deviations 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 behavior of $P(X(0) > \gamma)$ as γ gets large.

In the Section 3 we show how the large deviations approach can be applied in this situation. Section 4 gives the formal proof of our result.

2. Construction of a stationary solution

Our approach uses the classical representation of the steady state buffer content process. For each $t \geq 0$

$$X(t) = \max\left(\sup_{0<u\leq t}\int_u^t (N(s)-r)\,ds, X(0)+\int_0^t (N(s)-r)\,ds\right). \tag{2.1}$$

By the ergodic theorem, the finite dimensional distributions of $\{X(t+h), t \geq 0\}$ converge as $h \to \infty$ to the same limit for any $X(0)$ and any number and remaining lengths of the sessions present in the system at time 0. Therefore, we may and will assume that $X(0) = 0$ a.s. and that the process $\{N(t),\, t \geq 0\}$ is initialized as follows. The number of the sessions present in the system at time 0 has the Poisson distribution with mean equal to $\lambda\mu$ and, conditionally on $N(0) = k$, the remainders of the session lengths of the k sessions are independent with the common distribution distribution given by

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

Observe that if (1.2) holds, then

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

as a consequence of Karamata's theorem (Resnick (1987)). We denote by $N_1(t)$ the number of the $N(0)$ initial sessions still present in the system at time $t \geq 0$, and by $N_0(t)$ the number of sessions arriving in the interval $(0,t]$ and still running at time t. Clearly, the stochastic processes $\{N_0(t)\}$ and $\{N_1(t)\}$ are independent, and $N(t) = N_0(t) + N_1(t)$ for all $t \geq 0$. Note that the process $\{N_0(t),\, t \geq 0\}$ describes the number of customers in the system in a $M/G/\infty$ queue that starts empty at time 0. It is well known (see e.g. Kelly (1979)) and easy to check directly, that the stochastic process $\{N(t) = N_0(t) + N_1(t),\, t \geq 0\}$ is stationary.

With the system (1.1) initialized as above, the expression (2.1) for $X(t)$ becomes

$$X(t) = \sup_{0\leq u\leq t}\int_u^t (N(s)-r)\,ds, \quad t \geq 0.$$

Using the stationarity of the process $\{N(t),\, t \geq 0\}$ we immediately conclude that for every $h > 0$

$$\{X(t+h),\, t \geq 0\} \stackrel{\mathrm{d}}{=} \left\{\sup_{-t\leq u\leq h}\int_{-u}^t (N(s)-r)\,ds, \quad t \geq 0\right\} \tag{2.4}$$

in terms of equality of the finite dimensional distributions. Of course, the process $\{N(t),\, t \geq -h\}$ in the right hand side of (2.4) is, once again, stationary, and at time $-h$ it is initialized as described above. As $h \to \infty$ the right hand side of (2.4) converges with probability 1 for every $t \geq 0$, hence giving the stationary version of the buffer content process

$$\{X^*(t+h),\, t \geq 0\} \stackrel{\mathrm{d}}{=} \left\{ \sup_{u \geq -t} \int_{-u}^{t} (N(s) - r)\, ds, \quad t \geq 0 \right\}.$$

Since in this paper we are studying the stationary version of the buffer content process, we will lighten the notation (with some abuse of the latter) and work with the buffer content process $\{X(t),\, t \geq 0\}$ *defined* by

$$X(t) = \sup_{u \geq -t} \int_{-u}^{t} (N(s) - r)\, ds = \sup_{u \leq t} \int_{u}^{t} (N(s) - r)\, ds, \quad t \geq 0, \tag{2.5}$$

and $\{N(t),\, -\infty < t < \infty\}$ is stationary. That is, at any time s, $N(s)$ is Poisson with mean equal to $\lambda\mu$ and, conditionally on $N(s) = k$, the remainders of the session lengths of the k sessions present in the system at time s are iid with the common distribution distribution given by (2.2). Note that the processes $\{X(t),\, t \geq 0\}$ and $\{N(t),\, t \geq 0\}$ related by (2.5) still satisfy the equation (1.1). Furthermore, since the stationary process $\{N(t),\, t \geq 0\}$ (defined in the previous section) is reversible, we conclude that

$$X(0) \stackrel{\mathrm{d}}{=} \sup_{u \geq 0} \int_{0}^{u} (N(s) - r)\, ds. \tag{2.6}$$

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

For simplicity, we will take $X(0)$ as *defined* by the right hand side of (2.6). From now on we use the number of open sessions process $N(t)$ only for nonnegative times t, and hence we will use the notation introduced in Section 1: $N(t) = N_0(t) + N_1(t)$ for $t \geq 0$. Recall that $N_1(t)$ is the number of the $N(0)$ initial sessions still present in the system at time $t \geq 0$, and $N_0(t)$ is the number of sessions arriving in the interval $(0, t]$ and still running at time t.

Under the assumption (1.3) the stochastic process

$$S_u = \int_0^u (N(s) - r)\, ds =: A(u) - ru, \quad u \geq 0 \tag{3.1}$$

behaves similarly to a negative drift random walk, with an important difference, however. 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 $\{N(t),\, t \geq 0\}$ has only gradual effect. By effectively reducing the service rate while the session is running, the effect of such a session can be thought of as changing the relationship between the arrival rate and the service rate as experienced by the subsequently arriving sessions. In the present study we assume that

$$\lambda\mu < r < \lambda\mu + 1. \tag{3.2}$$

Since the assumption (1.3) is in force, the additional restriction provided by (3.2) is only $r < \lambda\mu + 1$. However, the meaning of this additional assumption is that, when a session is running, all additional sessions experience a supercritical queuing system, in which the server cannot cope with the offered traffic.

The event that the supremum of the negative drift process $\{S(u),\, u \geq 0\}$ in (3.1) 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 (1998) and Mikosch and Samorodnitsky (1999).

Under the assumption (3.2) the "least number of causes" mentioned above turns out to be equal to 1. This realization allows us to give a logically simple and transparent proof of our main result, Theorem 1 below. The large deviation approach does not obviate all technical difficulties but does provide a logical framework. There are certain technical difficulties involved in the proof that may obscure the rather simple idea, so we present an outline after stating the theorem.

Theorem 1. *Under the assumption of regular variation (1.2) and the assumption (3.2) we have*

$$P(X(0) > \gamma) \sim \frac{\lambda}{\alpha - 1} \frac{(\lambda\mu + 1 - r)^{\alpha - 1}}{r - \lambda\mu} \gamma^{-(\alpha-1)} L(\gamma) \tag{3.3}$$

as $\gamma \to \infty$.

Remark 2. This theorem was stated under somewhat stronger conditions 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 unproven fact nor the stronger assumption and, hence, should be attributed to them. A related work is in Likhanov and Mazumdar (1999). Note that the tail of the distribution of $X(0)$ is heavier than the tail of F.

3.1. **Informal justification.** Here is the promised outline of the argument. The reader needs to believe that under the assumption (3.2) one, and only one "cause" will force the stochastic process $\{S_u,\, u \geq 0\}$ in (3.1) to reach a high positive level γ. We expect that it is either one of the $N(0)$ sessions remaining in the system at time 0 that is long enough to cause the process $\{S_u,\, u \geq 0\}$ to reach level γ, or else the initial sessions do not contribute almost anything, and it is the newly arriving sessions that are counted in $\{N_0(t),\, t \geq 0\}$ that cause the system to get to level γ. In addition, if it is the initial sessions that make the process $\{S_u,\, u \geq 0\}$ reach level γ, then 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 $\{S_u,\, u \geq 0\}$ experiences a temporary positive drift of $\lambda\mu + 1 - r$, and so during this time the process $\{S_u,\, u \geq 0\}$ 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,

$$\begin{aligned} P(X(0) > \gamma) &= P(\sup_{u \geq 0} S_u > \gamma) \\ &\approx P\left(\text{one of the } N(0) \text{ initial sessions has remaining lifetime} > \frac{\gamma}{\lambda\mu + 1 - r}\right) \\ &\quad + P\left(\sup_{u \geq 0} \int_0^u (N_0(s) - r)\, ds > \gamma\right). \end{aligned} \tag{3.4}$$

However,

$$
\begin{aligned}
&P\Big(\text{one of the } N(0) \text{ initial sessions has remaining lifetime} > \frac{\gamma}{\lambda\mu+1-r}\Big)\\
&=\sum_{j=1}^{\infty} P(N(0)=j)P\left(\text{one of the } j \text{ initial sessions has remaining lifetime} > \frac{\gamma}{\lambda\mu+1-r}\right)\\
&\approx\sum_{j=1}^{\infty} P(N(0)=j)jP\left(\text{a generic initial session has remaining lifetime} > \frac{\gamma}{\lambda\mu+1-r}\right)
\end{aligned}
$$

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

$$
\begin{aligned}
&P\left(\text{one of the } N(0) \text{ initial sessions has remaining lifetime} > \frac{\gamma}{\lambda\mu+1-r}\right)\\
&\approx \left(1-F_*\left(\frac{\gamma}{\lambda\mu+1-r}\right)\right)E(N(0)) \qquad (3.5)\\
&\sim \frac{\lambda}{\alpha-1}(\lambda\mu+1-r)^{\alpha-1}\gamma^{-(\alpha-1)}L(\gamma)
\end{aligned}
$$

as $\gamma\to\infty$ by Karamata's theorem. See e.g. Resnick (1987).

On the other hand, the event

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

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,

$$
\begin{aligned}
&P\Big(\sup_{u\ge 0}\int_0^u (N_0(s)-r),\, ds>\gamma\Big)\\
&\approx P\Big(\text{there is a session of length } W \text{ arriving at time } T \text{ and such that}\\
&\qquad W(\lambda\mu+1-r)>\gamma+(r-\lambda\mu)T\Big)\,.
\end{aligned}
$$

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

$$
m(dt,dw)=\lambda\,dt\,F(dw) \qquad (3.6)
$$

we immediately see that

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

where the set A is defined by

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

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

$$P\left(\sup_{u\geq 0} \int_0^u (N_0(s) - r), ds > \gamma\right) \approx 1 - e^{-m(A)} \sim m(A) \tag{3.7}$$

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

as $\gamma \to \infty$.

Now one only has to substitute (3.5) and (3.7) into (3.4) to obtain (3.3).

We turn now to a formal proof.

4. Formal proof of Theorem 1

Proof of Theorem 1 We will prove that

$$\limsup_{\gamma\to\infty} \left(\gamma^{\alpha-1} L(\gamma)^{-1}\right) P(X(0) > \gamma) \leq \frac{\lambda}{\alpha - 1} \frac{(\lambda\mu + 1 - r)^{\alpha-1}}{r - \lambda\mu} . \tag{4.1}$$

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. We would like to mention, nevertheless, that the same approach we will use to prove (4.1) can also be used to prove the lower bound (and the technical details are quite a bit simpler). 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 M on $[0,\infty) \times \mathbb{R}_+$ by

$$\mathcal{M} = \sum_{i=1}^{N(0)} \epsilon_{(0,Y_i^*)} + \sum_{i=1}^{\infty} \epsilon_{(\Gamma_i,Y_i)}, \tag{4.2}$$

where

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

The random measure $\sum_{i=1}^{N(0)} \epsilon_{(0,Y_i^*)}$ is a Poisson process on $\{0\} \times \mathbb{R}_+$ and has mean measure $\delta_0(dt) \times F_*(dy)$ while $\sum_{i=1}^{\infty} \epsilon_{(\Gamma_i,Y_i)}$ 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

$$\begin{aligned} N_1(t) =& \sum_{i=1}^{N(0)} \epsilon_{(0,Y_i^*)}\Big(\{0\} \times (t,\infty)\Big) \\ =& \mathcal{M}\big(\{0\} \times (t,\infty)\big) \\ N_0(t) =& \mathcal{M}\big(\{(u,l) : 0 < u < t < u + l\}\big), \end{aligned}$$

and

$$N(t) = N_1(t) + N_2(t).$$

We start with preliminary separation of the effect of the initial sessions and that of subsequently arriving sessions as in the approximate equality in (3.4). We fix an $\epsilon \in (0,1)$ and write

$$\begin{aligned}
P(X(0) > \gamma) =& P(\sup_{u \geq 0} S_u > \gamma) \\
\leq& P\left(\bigvee_{u\geq 0} S_u > \gamma, \bigvee_{i=1}^{N(0)} Y_i^* > \epsilon\gamma\right) + P\left(\bigvee_{u\geq 0} S_u > \gamma, \bigvee_{i=1}^{N(0)} Y_i^* \leq \epsilon\gamma\right) \\
=:& P(A_{\gamma,\epsilon}) + P(B_{\gamma,\epsilon}).
\end{aligned} \tag{4.3}$$

We start with evaluation of the probability $P(A_{\gamma,\epsilon})$, which should be viewed as describing the effect of the $N(0)$ initial sessions. Let $Y^*_{(i)}$ for $i = 1, \dots, N(0)$ be the remaining lifetimes of the $N(0)$ initial sessions arranged in the 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 one. Recall that Y_i^*, $i \geq 1$ are iid 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 subexponential random variables in a Poisson sample are much less likely to be large simultaneously than just one (see Embrechts et al. (1979)). Write

$$P\left(Y^*_{(1)} > \epsilon\gamma, R > \epsilon\gamma\right) \leq P\left(Y^*_{(1)} > \epsilon\gamma, Y^*_{(2)} > \epsilon^2\gamma\right) + P\left(Y^*_{(1)} > \epsilon\gamma, N(0) > \epsilon^{-1}\right).$$

Observe that

$$\begin{aligned}
P\left(Y^*_{(1)} > \epsilon\gamma, Y^*_{(2)} > \epsilon^2\gamma\right) \leq& P[\sum_{i=1}^{N(0)} \epsilon_{Y_i^*}(\epsilon^2\gamma, \infty) \geq 2] \\
=& 1 - P[\sum_{i=1}^{N(0)} \epsilon_{Y_i^*}(\epsilon^2\gamma, \infty) = 0] - P[\sum_{i=1}^{N(0)} \epsilon_{Y_i^*}(\epsilon^2\gamma, \infty) = 1] \\
\leq& \frac{1}{2}\left(E\sum_{i=1}^{N(0)} \epsilon_{Y_i^*}(\epsilon^2\gamma, \infty)\right)^2 \\
=& \frac{1}{2}\left(\lambda\mu\bar{F}_*(\epsilon^2\gamma)\right)^2 \\
=& o(\bar{F}_*(\gamma)),
\end{aligned}$$

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

$$P\left(Y^*_{(1)} > \epsilon\gamma, Y^*_{(2)} > \epsilon^2\gamma\right) = o(P[Y_1^* > \gamma]).$$

Furthermore,

$$P[\bigvee_{i=1}^{N(0)} Y_i^* > \epsilon\gamma, N(0) > \epsilon^{-1}] = E1_{[N(0)>\epsilon^{-1}]} P[\bigvee_{i=1}^{N(0)} Y_i^* > \epsilon\gamma | N(0)]$$
$$\leq EN(0)1_{[N(0)>\epsilon^{-1}]}\bar{F}_*(\epsilon\gamma),$$

and thus we conclude

$$\limsup_{\gamma\to\infty} \frac{P\left(Y_{(1)}^* > \epsilon\gamma, R > \epsilon\gamma\right)}{P\left(Y_{(1)}^* > \gamma\right)} \leq \epsilon^{-(\alpha-1)} E\left(N(0)1_{[N(0)>\epsilon^{-1}]}\right)$$
$$\leq \epsilon^{-(\alpha-1)}\epsilon^k EN(0)^{k+1} \to 0 \tag{4.4}$$

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

$$P(A_{\gamma,\epsilon}) \leq P\left(A_{\gamma,\epsilon} \cap \{R \leq \epsilon\gamma\}\right) + P\left(Y_{(1)}^* > \epsilon\gamma, R > \epsilon\gamma\right). \tag{4.5}$$

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

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

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

$$\left\{\int_0^T n(s)\, ds + \sup_{S\leq u\leq T} \int_0^u (N(s) - n(s) - r)\, ds > \gamma\right\} \tag{4.6}$$

occurs as well for any $0 \leq S \leq T$. 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 $\{N(t),\, t \geq 0\}$. 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 (4.5) as follows. Let P_1 be a probability measure on the underlying space $(\Omega, \mathcal{F})$ under which the process $\{N(t),\, t \geq 0\}$ 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 $Y_1^* > \epsilon\gamma$. Using *Argument M* to bring in instantaneously all R units of work remaining in all initial sessions but the longest one, we conclude that

$$\begin{aligned}
&P\Big(A_{\gamma,\epsilon} \cap \{R \leq \epsilon\gamma\}\Big) \\
&\leq \sum_{n=0}^{\infty} e^{-\lambda\mu}\frac{(\lambda\mu)^n}{n!} n\, P(Y_1^* > \epsilon\gamma) P_1\left(\sup_{u\geq 0} S_u > (1-\epsilon)\gamma\right) \\
&\leq \lambda\mu\left[P_1\left(Y_1^* > \frac{\gamma(1-2\epsilon)}{\lambda\mu + 1 - r}\right) + P_1\left(\sup_{u\geq 0} S_u > (1-\epsilon)\gamma,\, Y_1^* \leq \frac{\gamma(1-2\epsilon)}{\lambda\mu + 1 - r}\right)\right] P[Y_1^* > \epsilon\gamma].
\end{aligned} \tag{4.7}$$

We will establish, as a consequence of the Law of Large Numbers, that

$$\lim_{\gamma\to\infty} P_1\left(\sup_{u\ge 0} S_u > (1-\epsilon)\gamma,\, Y_1^* \le \frac{(1-2\epsilon)}{\lambda\mu+1-r}\gamma\right) = 0\,. \tag{4.8}$$

Once (4.8) has been established, we may conclude by (4.4), (4.5), (4.7) and (4.8) that

$$\limsup_{\gamma\to\infty} \frac{P(A_{\gamma,\epsilon})}{\gamma^{-(\alpha-1)}L(\gamma)} \le \frac{\lambda}{\alpha-1}\left(\frac{1-2\epsilon}{\lambda\mu+1-r}\right)^{-(\alpha-1)} + \frac{\lambda}{\alpha-1}\epsilon^{-(\alpha-1)}E(N(0)1_{[N(0)>\epsilon^{-1}]})\,, \tag{4.9}$$

which is what we estimated in (3.5) (modulo letting $\epsilon\to 0$ which will be done later).

We now prove (4.8). For a $T>0$ we denote by

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

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

$$G(T) \ge \int_0^T N_0(s)\,ds \tag{4.10}$$

(since the right side only accounts for work accomplished by time T) and an immediate conclusion from the Strong Law of Large Numbers that

$$\frac{G(T)}{T} \to \lambda\mu \quad \text{a.s. as } T\to\infty\,. \tag{4.11}$$

By (4.11)

$$\begin{aligned}
&P_1\Big(\sup_{u\ge 0} S_u > (1-\epsilon)\gamma,\, Y_1^* \le \frac{(1-2\epsilon)}{\lambda\mu+1-r}\gamma\Big)\\
&\quad= P_1\left(\sup_{u\ge 0} S_u > (1-\epsilon)\gamma,\, Y_1^* \le \frac{\gamma(1-2\epsilon)}{\lambda\mu+1-r},\, G(T) \le \left(\frac{1-\epsilon}{1-2\epsilon}\right)^{1/2} T\lambda\mu\ \forall T>\gamma^{1/3}\right) + o(1)\\
&\quad:= P_1\left(A^{(1)}_{\gamma,\epsilon}\right) + o(1)
\end{aligned}$$

as $\gamma\to\infty$.

We can bound $P_1\left(A^{(1)}_{\gamma,\epsilon}\right)$ in the following manner. Write

$$\begin{aligned}
P_1\left(A^{(1)}_{\gamma,\epsilon}\right) \le & P_1\Bigg(\sup_{\gamma^{1/3}\le u\le Y_1^*} S_u > (1-\epsilon)\gamma,\, Y_1^* \le \frac{(1-2\epsilon)}{\lambda\mu+1-r}\gamma, \\
& \qquad G(T) \le \left(\frac{1-\epsilon}{1-2\epsilon}\right)^{1/2} T\lambda\mu,\ \forall T > \gamma^{1/3}\Bigg) \\
& + P_1\Bigg(\sup_{u\ge Y_1^*} S_u > (1-\epsilon)\gamma,\, Y_1^* \le \frac{(1-2\epsilon)}{\lambda\mu+1-r}\gamma, \\
& \qquad G(T) \le \left(\frac{1-\epsilon}{1-2\epsilon}\right)^{1/2} T\lambda\mu,\ \forall T > \gamma^{1/3}\Bigg) \\
& + P\left(N_0((0,\gamma^{1/3}]) \ge (1-\epsilon)\gamma^{2/3}-1\right) \\
:= & P_1\left(A^{(2)}_{\gamma,\epsilon}\right) + P_1\left(A^{(3)}_{\gamma,\epsilon}\right) + P_1\left(A^{(4)}_{\gamma,\epsilon}\right),
\end{aligned}$$

where we use the notation $N_0(B)$ for the number of sessions initiated in a (Borel) set $B\subset(0,\infty)$. The term $P_1\left(A^{(4)}_{\gamma,\epsilon}\right)$ results from the fact that if $u<\gamma^{1/3}$ and $N_0(0,\gamma^{1/3}]\le(1-\epsilon)\gamma^{2/3}-1$, then

$$\begin{aligned}
S_u = & \int_0^u (N(s)-r)ds \\
= & (1-r)u + \int_0^u N_0(s)ds \\
\le & (1-r)u + \int_0^u \left((1-\epsilon)\gamma^{2/3}-1\right)ds \\
\le & (1-r)\gamma^{1/3} + \gamma^{1/3}[(1-\epsilon)\gamma^{2/3}-1] \\
= & \gamma^{1/3}(-r) + (1-\epsilon)\gamma \\
< & (1-\epsilon)\gamma.
\end{aligned}$$

Note that $P_1\left(A^{(4)}_{\gamma,\epsilon}\right) = o(1)$, as $\gamma\to\infty$.

Observe that by (4.10) for P_1- almost every $\omega\in A^{(2)}_{\gamma,\epsilon}$ and $\gamma^{1/3}\le u\le Y_1^*$,

$$\begin{aligned}
S_u = & \int_0^u (N(s)-r)\,ds \le (1-r)u + \int_0^u N_0(s)\,ds \le (1-r)u + G(u) \\
\le & (1-r)u + \left(\frac{1-\epsilon}{1-2\epsilon}\right)^{1/2} u\lambda\mu \le u\left(\lambda\mu+1-r\right)\left(\frac{1-\epsilon}{1-2\epsilon}\right)^{1/2} \\
\le & \gamma\Big((1-\epsilon)(1-2\epsilon)\Big)^{1/2} < (1-\epsilon)\gamma\,.
\end{aligned}$$

Hence $P_1\left(A^{(2)}_{\gamma,\epsilon}\right) = 0$, for all $\gamma>0$.

It remains to evaluate the probability $P_1(A^{(3)}_{\gamma,\epsilon})$. To do so we use, once again, *Argument M.* We bring in instantaneously at time 0 all the work in the only initial session of length Y_1^* and all the work in sessions arriving in time interval $(0, Y_1^*]$. For P_1–almost every $\omega \in A^{(3)}_{\gamma,\epsilon}$ this amounts to bringing in

$$G(Y_1^*) + Y_1^* \le Y_1^* \left(\frac{1-\epsilon}{1-2\epsilon}\right)^{1/2} \lambda\mu + Y_1^*$$

units of work. We conclude that

$$\begin{aligned} P_1(A^{(3)}_{\gamma,\epsilon}) \le & P_1\Bigg(\sup_{u\ge 0}\int_0^u (N_0(s)-r)\,ds > (1-\epsilon)\gamma - \left(\frac{1-\epsilon}{1-2\epsilon}\right)^{1/2} \lambda\mu Y_1^* + Y_1^* - rY_1^*, \\ & \qquad Y_1^* \le \frac{(1-2\epsilon)}{\lambda\mu+1-r}\gamma\Bigg) \\ \le & P_1\Bigg(\sup_{u\ge 0}\int_0^u (N_0(s)-r)\,ds > (1-\epsilon)\gamma - Y_1^*\left(\frac{1-\epsilon}{1-2\epsilon}\right)^{1/2} (\lambda\mu+1-r), \\ & \qquad Y_1^* \le \frac{(1-2\epsilon)}{\lambda\mu+1-r}\gamma\Bigg) \\ \le & P_1\Bigg(\sup_{u\ge 0}\int_0^u (N_0(s)-r)\,ds > \gamma\Big((1-\epsilon) - ((1-\epsilon)(1-2\epsilon))^{1/2}\Big)\Bigg) = o(1) \end{aligned}$$

as $\gamma \to \infty$ because the negative drift condition (1.3) requires

$$P_1[\sup_{u\ge 0}\int_0^u (N_0(s)-r)\,ds < \infty] = 1.$$

Therefore, we have proved (4.8), and so (4.9) follows.

We continue with evaluation of the probability $P(B_{\gamma,\epsilon})$ in (4.3), and it should be viewed as describing the effect of the sessions initiated after time 0, as in (3.7). 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,

$$P\left(Y^*_{(1)} + R > \gamma, Y^*_{(1)} \le \gamma\right) = o(P(Y^*_{(1)} > \gamma)),$$

and hence

$$
\begin{aligned}
(4.12) \qquad P(B_{\gamma,\epsilon}) =& P[\sup_{u\ge 0} S_u > \gamma, \bigvee_{i=1}^{N(0)} Y_i^* \le \epsilon\gamma] \\
\le& P[Y_{(1)}^* + R + \sup_{u\ge 0}\int_0^u (N_0(s)-r)ds > \gamma, \bigvee_{i=1}^{N(0)} Y_i^* \le \epsilon\gamma] \\
\le& P[Y_{(1)}^* + R > \epsilon\gamma, Y_{(1)}^* \le \epsilon\gamma] + P[\sup_{u\ge 0}\int_0^u (N_0(s)-r)ds > (1-\epsilon)\gamma] \\
=& P[\sup_{u\ge 0}\int_0^u (N_0(s)-r)\,ds > (1-\epsilon)\gamma] + o(\gamma^{-(\alpha-1)}L(\gamma)),
\end{aligned}
$$

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

$$
(4.13) \qquad \limsup_{\gamma\to\infty} \left(\gamma^{\alpha-1}L(\gamma)^{-1}\right) P\left(\sup_{u\ge 0}\int_0^u (N_0(s)-r)\,ds > \gamma\right) \le \frac{\lambda}{\alpha-1}\frac{(\lambda\mu+1-r)^\alpha}{r-\lambda\mu}\,.
$$

Assuming that (4.13) has been proved, we will have by (4.12) and (4.13)

$$
\limsup_{\gamma\to\infty} \left(\gamma^{\alpha-1}L(\gamma)^{-1}\right) P(B_{\gamma,\epsilon}) \le (1-\epsilon)^{-(\alpha-1)}\frac{\lambda}{\alpha-1}\frac{(\lambda\mu+1-r)^\alpha}{r-\lambda\mu}\,,
$$

which together with (4.3) and (4.9) shows that

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

and so our claim (4.3) follows by letting $\epsilon \to 0$.

Therefore, it remains to prove (4.13). 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 $\{\sup_{u\ge 0}\int_0^u (N_0(s)-r)\,ds > \gamma\}$ 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 $\{\int_0^u (N_0(s)-r)\,ds,\ u\ge 0\}$ decays, roughly, linearly. Formally,

Lemma 1. *We have*

$$
(4.14) \qquad \lim_{M\to\infty}\limsup_{\gamma\to\infty} \left(\gamma^{\alpha-1}L(\gamma)^{-1}\right) P\left(\sup_{u\ge M\gamma}\int_0^u (N_0(s)-r)\,ds > \gamma\right) = 0\,.
$$

The next lemma formally establishes that for the event $\{\sup_{u\ge 0}\int_0^u (N_0(s)-r)\,ds > \gamma\}$ to occur, at least one very long session has to be initiated by the time of crossing of level γ. It has the same spirit as Lemma 2.7 in Mikosch and Samorodnitsky (1998). Recall the random measure $\mathcal{M}$ defined in (4.2).

Lemma 2. *For every $M > 0$ there is an $\epsilon_0 > 0$ such that for all $0 < \epsilon < \epsilon_0$*

$$
(4.15) \qquad \lim_{\gamma\to\infty}\frac{P\big(\sup_{0\le u\le M\gamma}\int_0^u (N_0(s)-r)\,ds > \gamma, \mathcal{M}\big((0,M\gamma]\times(\epsilon\gamma,\infty)\big)=0\big)}{\gamma^{-(\alpha-1)}L(\gamma)} = 0
$$

and

$$\lim_{\gamma\to\infty} \frac{P\big(\sup_{0\le u\le M\gamma}(G(u)-ru)>\gamma, \mathcal{M}\big((0,M\gamma]\times(\epsilon\gamma,\infty)\big)=0\big)}{\gamma^{-(\alpha-1)}L(\gamma)}=0. \tag{4.16}$$

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

$$A(\epsilon,M;\gamma)=\Big\{\sup_{0\le u\le M\gamma}\int_0^u (N_0(s)-r)\,ds>\gamma, \mathcal{M}\Big((0,M\gamma]\times(\epsilon\gamma,\infty)\Big)=1\Big\}.$$

So $A(\epsilon,M;\gamma)$ is the event of a crossing of level γ by time $M\gamma$ while exactly one session length exceeds $\epsilon\gamma$. Since

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

as $\gamma\to\infty$ for all fixed M and ϵ, it follows from Lemmas 1 and 2 that (4.13) will follow once we show that (for any fixed $M\ge 1$)

$$\limsup_{\gamma\to\infty}\left(\gamma^{\alpha-1}L(\gamma)^{-1}\right)P\left(A(\epsilon,M;\gamma)\right)\le K_\epsilon\frac{\lambda}{\alpha-1}\frac{(\lambda\mu+1-r)^\alpha}{r-\lambda\mu} \tag{4.17}$$

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

Let T_* and T^* be the times when the first session of length exceeding $\epsilon\gamma$ is initiated and is finished, correspondingly. Observe that T_* and T^* are well defined random variables, and that on the event $A(\epsilon,M;\gamma)$ they are also the times of initiation and completion of the only session of length exceeding $\epsilon\gamma$ that is initiated in $(0,M\gamma]$. For a $\delta\in(0,1)$ we split the event $A(\epsilon,M;\gamma)$ into two, depending on whether or not the level γ was first exceeded by the time T^* or not. If not, we use also *Argument M* and bring instantaneously at time T^* all work remaining in sessions running at that time. Define

$$\begin{aligned}A_1(\epsilon,M;\gamma)=&\Big\{\Big(\Big[\sup_{0\le u\le T^*\wedge M\gamma}\int_0^u (N_0(s)-r)\,ds>\gamma\Big]\cup\Big[G(T^*\wedge M\gamma)-r(T^*\wedge M\gamma)>\gamma\Big]\Big)\\ &\qquad\cap\Big[\mathcal{M}\big((0,M\gamma]\times(\epsilon\gamma,\infty)\big)=1\Big]\Big\},\end{aligned}$$

and we get

$$\begin{aligned}P\big(A(\epsilon,M;\gamma)\big)\le& P\Big(\sup_{0\le u\le M\gamma}\int_0^u (N_0(s)-r)\,ds>\delta\gamma, \mathcal{M}\big((0,M\gamma]\times(\epsilon\gamma,\infty)\big)=0\Big)\\ &+P\left(A_1(\epsilon,M/(1-\delta);(1-\delta)\gamma)\right).\end{aligned} \tag{4.18}$$

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 in the right hand side of (4.18) is of an order smaller than $\gamma^{-(\alpha-1)}L(\gamma)$ if ϵ is small enough relatively to δ. Therefore, (4.17) will follow once we show that for any fixed $M>1$,

$$\limsup_{\gamma\to\infty}\left(\gamma^{\alpha-1}L(\gamma)^{-1}\right)P\left(A_1(\epsilon,M;\gamma)\right)\le K_\epsilon\frac{\lambda}{\alpha-1}\frac{(\lambda\mu+1-r)^\alpha}{r-\lambda\mu} \tag{4.19}$$

for some $K_\epsilon\to 1$ as $\epsilon\to 0$ (we are abusing the notation a bit by using the same K_ϵ in both (4.17) and (4.19)).

We further split the event $A_1(\epsilon,M;\gamma)$ depending on whether or not the level γ was first exceeded by the time T_* or not. If not, we use once again *Argument M* and bring instantaneously at time T_* all work remaining in sessions running at that time. Observe that on the event $A_1(\epsilon,M;\gamma)$ by the time $T_* \wedge M\gamma$ only sessions of length not exceeding $\epsilon\gamma$ are initiated, and, apart from the only long session, only sessions of length not exceeding $\epsilon\gamma$ are initiated between $T_* \wedge M\gamma$ and $M\gamma$. To make the accounting easier, we define a new process, say, $\{N_0^{(1)}(t),\, t \geq 0\}$ 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 $\lambda F(\epsilon\gamma)$ and the distribution of the job lengths is given by

$$F^{(1)}(A) = \frac{F\left(A \cap [0, \epsilon\gamma]\right)}{F([0, \epsilon\gamma])} \quad A \text{ Borel}.$$

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 $(\Omega, \mathcal{F}, P)$. We have

$$P\left(A_1(\epsilon, M; \gamma)\right) \leq P\left(A_2(\epsilon, M; \gamma)\right) + P\left(A_3(\epsilon, M; \gamma)\right), \tag{4.20}$$

where

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

and

$$A_3(\epsilon, M; \gamma) = \Big\{ T^* \leq M\gamma \text{ and either } G^{(1)}(T_*) - rT_* + \sup_{0 \leq u \leq T^* \wedge M\gamma - T_* \wedge M\gamma} \int_0^u (N_0^{(2)}(s) + 1 - r)\, ds > \gamma,$$

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

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

Clearly, Lemmas 1 and 2 imply that the first term in the right hand of (4.20) is of an order smaller than $\gamma^{-(\alpha-1)} L(\gamma)$ if ϵ is small enough. Therefore, (4.19) will be proved if we show that for all $M > 1$

$$\limsup_{\gamma \to \infty} \left(\gamma^{\alpha-1} L(\gamma)^{-1}\right) P\left(A_3(\epsilon, M; \gamma)\right) \leq K_\epsilon \frac{\lambda}{\alpha - 1} \frac{(\lambda\mu + 1 - r)^\alpha}{r - \lambda\mu} \tag{4.21}$$

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

We have by (4.10) for any $\delta > 0$

$$P\left(A_3(\epsilon, M; \gamma)\right) \leq 2P\left(G^{(1)}(t) - (\lambda\mu + \frac{\delta}{2M})t > \frac{\delta}{2}\gamma \text{ for some } 0 \leq t \leq M\gamma\right) \tag{4.22}$$

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

By Lemma 2 we have

$$\lim_{\gamma \to \infty} \left(\gamma^{\alpha-1} L(\gamma)^{-1}\right) P\left(G^{(1)}(t) - (\lambda\mu + \frac{\delta}{2M})t > \frac{\delta}{2}\gamma \text{ for some } 0 \leq t \leq M\gamma\right) = 0 \tag{4.23}$$

as long as ϵ is small enough relative to δ, and so we only have to treat the second term in the right hand side of (4.22), which is exactly the term corresponding to the main probability term which we estimated in (3.7).

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

$$F^{(2)}(A)=\frac{F\left(A\cap(\epsilon\gamma,\infty)\right)}{F((\epsilon\gamma,\infty))},\quad A \text{ Borel}.$$

Observe that

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

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

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

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 (4.21) will follow if for every fixed $M>1$

(4.24)
$$\limsup_{\gamma\to\infty}\frac{P[(\lambda\mu+1-r)(T^*-T_*)+(\lambda\mu-r)T_*>\gamma,\ T^*\le M\gamma]}{\gamma^{-(\alpha-1)}L(\gamma)}\le K_\epsilon\frac{\lambda}{\alpha-1}\frac{(\lambda\mu+1-r)^\alpha}{r-\lambda\mu},$$

once again, for some $K_\epsilon\to1$ as $\epsilon\to0$.

However, denoting

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

and using regular variation together with the dominated convergence theorem we see that for every $0<\epsilon<1$

$$\begin{aligned}P\Big((\lambda\mu+1-r)(T^*-T_*)+&(\lambda\mu-r)T_*>\gamma,\ T^*\le M\gamma\Big)\\ &\le\frac{1}{1-F(\epsilon\gamma)}\int_0^\infty e^{-t}\left(1-F\left(\frac{\gamma+(r-\lambda\mu)\frac{t}{\lambda(1-F(\epsilon\gamma))}}{\lambda\mu+1-r}\right)\right)dt\\ &=\gamma\int_0^\infty \exp\{-\gamma(1-F(\epsilon\gamma))\}\left(1-F\left(\gamma(a+bt)\right)\right)dt\\ &\sim\gamma(1-F(\gamma))\int_0^\infty(a+bt)^{-\alpha}\,dt\\ &=\frac{a^{-(\alpha-1)}}{(\alpha-1)b}\gamma(1-F(\gamma))\\ &=\frac{\lambda}{\alpha-1}\frac{(\lambda\mu+1-r)^\alpha}{r-\lambda\mu}\gamma(1-F(\gamma)),\end{aligned}$$

thus proving (4.24) and, hence, completing the proof of the theorem. □

5. Proofs of Lemmas 1 and 2

The idea under the argument in both cases is to compare the fluid queue to an appropriate random walk with a negative drift, which crosses a positive level before the fluid queue does (see e.g. (4.10)).

Proof of Lemma 1 We have by (4.10) for all $u \geq M\gamma$

$$\int_0^u (N_0(s)-r)\,ds = \int_0^{M\gamma} (N_0(s)-r)\,ds + \int_{M\gamma}^u (N_0(s)-r)\,ds \leq G(M\gamma) - rM\gamma + \int_{M\gamma}^u (N_0(s)-r)\,ds\,,$$

and so

(5.1)

$$P\Big(\sup_{u\geq M\gamma}\int_0^u (N_0(s)-r)\,ds > \gamma\Big)$$
$$\leq P\Big(\sum_{j=1}^K Y_j - rM\gamma > -\frac{r-\lambda\mu}{2}M\gamma\Big) + P\Big(\sup_{u\geq 0}\int_0^u (N_0(s)-r)\,ds > \gamma\Big(1+\frac{r-\lambda\mu}{2}M\Big)\Big),$$

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

Observe that by (4.10)

$$\sup_{u\geq 0}\int_0^u (N_0(s)-r)\,ds \leq \sup_{n\geq 0} S_n\,,$$

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

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

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

$$\lim_{M\to\infty}\limsup_{\gamma\to\infty}\left(\gamma^{\alpha-1}L(\gamma)^{-1}\right)P\left(\sup_{u\geq 0}\int_0^u (N_0(s)-r)\,ds > \gamma\left(1+\frac{r-\lambda\mu}{2}M\right)\right) = 0,. \tag{5.2}$$

Furthermore,

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

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

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

as $\gamma \to \infty$ for some $c > 0$. On the other hand,

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

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

$$\lim_{M\to\infty} \limsup_{\gamma\to\infty} \left(\gamma^{\alpha-1} L(\gamma)^{-1}\right) P\left(\sum_{j=1}^{K} Y_j - rM\gamma > -\frac{r-\lambda\mu}{2} M\gamma\right) = 0\,, \tag{5.3}$$

and now (4.14) follows from (5.2) and (5.3).

This completes the proof of Lemma 1. □

Proof of Lemma 2 By (4.10) the statement (4.15) follows from the statement (4.16), and so it is enough to prove the latter. Clearly,

$$P\left(\sup_{0\le u\le M\gamma} (G(u) - ru) > \gamma, \mathcal{M}\big((0, M\gamma] \times (\epsilon\gamma, \infty)\big) = 0\right) \le P\left(\sup_{n\ge k(\epsilon\gamma)} S_n > \gamma\right),$$

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

$$k(\epsilon\gamma) = \inf\{n \ge 0 :\ Z_n > \epsilon\gamma\}\,.$$

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

$$P\left(\sup_{n\ge k(\epsilon\gamma)} S_n > \gamma\right) \le P\left(\sup_{0\le n\le\gamma/M} S_n > \gamma\right) + P\left(\sup_{\gamma/M\le n\le k(\epsilon\gamma)} S_n > \gamma\right)$$
$$:= p_1(M;\gamma) + p_2(\epsilon, M;\gamma)\,.$$

However, by Lemma 2.5 of Mikosch and Samorodnitsky (1998)

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

while by Lemma 2.7 of Mikosch and Samorodnitsky (1998), for every fixed $M > 1$ and for all ϵ small enough,

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

Therefore, the statement (4.16) follows, and the proof is complete. □

References

R. Agrawal, A. Makowski and P. Nain (1998): On a reduced load equivalence for fluid queues under subexponentiality. Preprint.

M. Arlitt and C. Williamson (1996): Web servers workload characterization: The search for invariants (extended version). In *Proceedings of the ACM Sigmetrics Conference, Philadelphia, PA*.

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

J. Beran, R. Sherman, W. Willinger and M. Taqqu (1995): Long-range dependence in Variable-bit-rate video traffic. *IEEE Transactions on Communications* 43:1566–1579.

O. Boxma (1996): Regular variation in a multi–source fluid queue. Technical Report BS–R9614, CWI.

O. Boxma and V. Dumas (1996): Fluid queues with long–tailed activity period distributions. To appear in a Special Issue of *Stochastic Analysis and Optimization of Communication Systems*.

G. L. Choudhury and W. Whitt (1995): Long-tail buffer-content distributions in broadband networks. Preprint.

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

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

M. Crovella and A. Bestavros (1996): Self–similarity in World Wide Web traffic: evidence and possible causes. *Performance Evaluation Review* 24:160–169.

C. Cunha, A. Bestavros and M. Crovella (1995): Characteristics of www client–based traces. Preprint available as BU-CS-95-010 from {crovella,best}cs.bu.edu.

P. Embrechts, C. Goldie and N. Veraverbeke (1979): Subexponentiality and infinite divisibility. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 49:335–347.

P. Embrechts and N. Veraverbeke (1982): Estimates for the probability of ruin with special emphasis on the possibility of large claims. *Insurance: Mathematics and Economics* 1:55–72.

D. Heath, S. Resnick and G. Samorodnitsky (1997a): How system performance is affected by the interplay of averages in a fluid queue with long range dependence induced by heavy tails. *Annals of Applied Probability*, to appear. Available as hit-bounds.ps at http://www.orie.cornell.edu/ gennady/techreports.

D. Heath, S. Resnick and G. Samorodnitsky (1997b): Patterns of buffer overflow in a class of queues with long memory in the input stream. *The Annals of Applied Probability* 7:1021–1057.

D. Heath, S. Resnick and G. Samorodnitsky (1998): Heavy tails and long range dependence in on/off processes and associated fluid models. *Mathematics of Operations Research* 23:145–165.

P. Jelenković and A. Lazar (1999): Asymptotic results for multiplexing subexponential on–off sources. *Advances in Applied Probability* 31:...

F. Kelly (1979): *Reversibility and Stochastic Networks*. Wiley and Sons, Chichester.

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

N. Likhanov and R. Mazumdar (1999): Forthcoming.

Z. Liu, P. Nain, D. Towsley and Z.-L. Zhang (1997): Asymptotic behavior of a multiplexer fed by a long–range dependent process. Technical Report CMPSRI 97-16, University of Massachusetts at Amherst.

T. Mikosch and G. Samorodnitsky (1998): The supremum of a negative drift random walk with dependent heavy–tailed steps. Preprint. Available as walk.ps at www.orie.cornell.edu/ gennady/techreports.

T. Mikosch and G. Samorodnitsky (1999): Ruin probability with claims modeled by a stationary ergodic stable process. Preprint. Available as rstab.ps at www.orie.cornell.edu/ gennady/techreports.

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

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

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

S. Resnick and H. Rootzén (1998): Self–similar communication models and very heavy tails. Technical Report NO 1998-52, Department of Mathematics, Chalmers University of Technology – Götebrg University.

S. Resnick and G. Samorodnitsky (1997): Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues. *QUESTA*, to appear. Available as inf-server.ps at www.orie.cornell.edu/ gennady/techreports.

S. Resnick and G. Samorodnitsky (1998): Fluid Queues, Leaky Buckets, On-Off Processes and Teletraffic Modeling with Highly Variable and Correlated Inputs. *SELF-SIMILAR NETWORK TRAFFIC AND PERFORMANCE EVALUATION*, (Kihong Park and Walter Willinger, Editors; Wiley) to appear. Available as wiley.ps at www.orie.cornell.edu/ gennady/techreports.

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

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

A. Zwart (1998): A fluid queue with a finite buffer and subexponential input. Preprint.

School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853

E-mail address: `sid@orie.cornell.edu` `gennady@orie.cornell.edu`