HOW SYSTEM PERFORMANCE IS AFFECTED BY THE INTERPLAY OF AVERAGES IN A FLUID QUEUE WITH LONG RANGE DEPENDENCE INDUCED BY HEAVY TAILS

DAVID HEATH, SIDNEY RESNICK, AND GENNADY SAMORODNITSKY

Abstract. We consider a fluid queue with sessions arriving according to a Poisson process. A long–tailed distribution of session lengths induces long range dependence in the system and causes its performance to deteriorate. The deterioration is due to occurrence of load regimes far from average ones. Nonetheless, the *extent* of this performance deterioration is shown to depend crucially on the average values of the system parameters.

1. Introduction

We consider the following fluid queuing model. Sessions (ON periods) are initiated at a network server or multiplexer according to a Poisson process with rate $\lambda > 0$. Each session is active for a random length of time with distribution F and a finite mean μ; during this time it generates network traffic at unit rate. We assume that the lengths of different sessions are independent, and they are also independent of the Poisson arrival process. The service rate is $r > 0$ units of traffic per unit time. If $X(t)$ denotes the amount of work (measured in units of network traffic, e.g. packets) in the buffer at time t, then the content process $\{X(t),\, t \geq 0\}$ satisfies

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

where $N(t)$ is the number of sessions active at time t. Observe that $\{N(t),\, t \geq 0\}$ is the stochastic process describing the number of customers in the system in an $M/G/\infty$ queue. The mean values of the system parameters determine whether the system described by (1.1) reaches steady state, and we will always assume that

$$r > \lambda\mu. \tag{1.2}$$

That is, the maximal service rate in the system exceeds the overall arrival rate in the system.

The performance measure we are interested in is the expected time until overflow of a large buffer. Specifically, assuming that the work in the system is "collected" in a buffer of size γ, we call

$$\tau_\gamma = \inf\{t \geq 0 :\, X(t) \geq \gamma\} \tag{1.3}$$

the time of buffer overflow. A system has good performance if the expected time of buffer overflow $E\tau_\gamma$ is large. The tail behavior of the session length distribution F has crucial impact on the rate of growth of the expected time of buffer overflow as $\gamma \to \infty$. Exponentially fast decaying

---

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

Key words and phrases. fluid queue, heavy tails, long range dependence, performance of a queue, $M/G/\infty$ queue, time until overflow, large deviations, association.

Resnick's and Samorodnitsky's research was partially supported by NSF Grants DMS-97-04982 and DMI-9713549 at Cornell University.

probability tails of F have been found to imply in similar situations exponentially fast increase in $E\tau_\gamma$ as a function of γ. See, e.g., Heath et al. (1997). On the other hand, power-like decay of probability tails of F have been shown, in certain circumstances, to lead to a polynomially fast increase in $E\tau_\gamma$ as a function of γ (Resnick and Samorodnitsky (1997a)), and a similar phenomenon has been observed in Heath et al. (1997).

It has been argued in the literature that a decay in system performance is caused by long range dependence in the input stream. This has been observed in different situations by Duffield and O'Connell (1995), Ryu and Lowen (1995), Erramilli et al. (1996), Heath et al. (1996), Vamvakos and Anantharam (1996), Liu et al. (1997) and Resnick and Samorodnitsky (1997b) A survey is in Boxma and Dumas (1996). Since heavy tailed session length is known to cause long range dependence in our model, and in similar models (Jelenković and Lazar (1995), Boxma and Dumas (1996), Heath et al. (1996), Willinger et al. (1996), Resnick and Samorodnitsky (1997b)), the loss in performance of our fluid queue we mentioned above is not surprising. It is also not surprising that the performance loss tends to grow as the session length distribution tails grow heavier, because the length of memory tends to increase with heaviness of the distribution tails. What is surprising is that the extent of the performance loss is determined by an interesting interplay of the heaviness of the tails and the average characteristics of the system, as described presently. Let

$$k = \inf\{j \geq 1 : \ \lambda\mu + j - r > 0\}. \tag{1.4}$$

Clearly, the parameter k is determined by the average characteristics of the system. It is the minimal number of sessions, running simultaneously, required to change the direction of the drift in the system from negative to positive. It turns out from the nature of large deviations in the heavy tailed situation, that this is exactly when the amount of work in the buffer goes up by any significant amount. Therefore, if the buffer is large, the buffer overflow will most likely occur when at least k sessions are running simultaneously, and for a sufficiently long period of time. Since the time until this occurs is, clearly, significantly affected by the value of k, the importance of the parameter k for system performance becomes less surprising. We refer the reader to Embrechts et al. (1997) and Resnick and Samorodnitsky (1997a) for a discussion of heavy tailed large deviations.

Various assumptions have been used in the literature to model heavy session lengths. In this paper we will impose only relatively weak assumptions. Specifically, we assume that the session length distribution tail $\overline{F} = 1 - F$ is *dominatedly varying.* That is,

$$\varliminf_{x\to\infty} \frac{\overline{F}(ax)}{\overline{F}(x)} > 0 \tag{1.5}$$

for some (equivalently, all) $a > 1$. Recall that such a distribution has finite *Matuszewska indices* $\infty > \alpha > \beta \geq 0$ such that for every $\epsilon > 0$ there are C and x_0 such that

$$\frac{1}{C} a^{-\alpha-\epsilon} \leq \frac{\overline{F}(ax)}{\overline{F}(x)} \leq C a^{-\beta+\epsilon} \tag{1.6}$$

for all $a > 1$ and $x \geq x_0$. See (Bingham et al., 1987, page 65, 71). To guarantee existence of a finite first moment we will often assume that $\beta > 1$.

2. Preliminary results

In this section we collect auxiliary results that are needed in this paper. These results deal with various aspects of the fluid model (1.1) and the underlying $M/G/\infty$ queue. We are especially interested in positive dependence occurring in these models. Our first result is a version of the standard exponential one–sided large deviation bound.

Lemma 1. *Let $\{F_\gamma,\, \gamma > 0\}$ be a uniformly integrable family of probability distributions on $[0,\infty)$ such that*

$$\int_0^\infty x F_\gamma\,(dx) \to \mu > 0$$

as $\gamma \to \infty$. For a $\gamma > 0$ let $\{Z_i^{(\gamma)},\, i \geq 1\}$ be a sequence of i.i.d. random variables with common distribution F_γ, independent of a Poisson process $\{N_\gamma(t),\, t \geq 0\}$ with intensity λ_γ such that

$$\lim_{\gamma\to\infty} \lambda_\gamma = \lambda > 0.$$

Denote

$$W_\gamma = \sum_{i=1}^{N_\gamma(\gamma)} Z_i^{(\gamma)}, \quad \gamma > 0.$$

Then for every $0 < \rho < 1$ there is a $c_\rho > 0$ such that

$$P(W_\gamma \leq \rho\lambda\mu\gamma) = o(e^{-c_\rho\gamma}) \tag{2.1}$$

as $\gamma \to \infty$.

Proof. We give a short proof for completeness. Exponentiating and using Markov's inequality gives us for any $\theta > 0$

$$P(W_\gamma \leq \rho\lambda\mu\gamma) \leq e^{\theta\rho\lambda\mu\gamma} E e^{-\theta W_\gamma} = \exp\Big\{\theta\rho\lambda\mu\gamma - \lambda_\gamma\gamma \int_0^\infty (1 - e^{-\theta x})\, F_\gamma(dx)\Big\}.$$

Consider now only γ so large that $\lambda_\gamma > \rho^{.25}\lambda$ and $\int_0^\infty x F_\gamma(dx) > \rho^{.25}\mu$. By the uniform integrability we can choose M so big that $\int_M^\infty x F_\gamma(dx) \leq (\rho^{.25} - \rho^{.4})\mu$ for all $\gamma > 0$. Finally, choose a θ so close to zero that $1 - e^{-\theta x} \geq \rho^{.25}\theta x$ for all $0 < x \leq M$. Then

$$\begin{aligned} P(W_\gamma \leq \rho\lambda\mu\gamma) &\leq \exp\Big\{\theta\rho\lambda\mu\gamma - \theta\rho^{.5}\lambda\gamma \int_0^M x F_\gamma(dx)\Big\} \\ &\leq \exp\Big\{\theta\rho\lambda\mu\gamma - \theta\rho^{.5}\lambda\gamma(\rho^{.25}\mu - (\rho^{.25} - \rho^{.4})\mu)\Big\} \\ &= \exp\{-(\rho^{.9} - \rho)\theta\lambda\mu\gamma\}, \end{aligned}$$

and we obtain (2.1) with any $c_\rho < (\rho^{.9} - \rho)\theta\lambda\mu$. □

The next lemma treats level crossings of a certain Markov chain with a negative drift. The c's in this lemma do not have to be the same. In general, we reserve the letter c for a finite positive constant whose value is immaterial and which may change each time it appears.

Lemma 2. *Let X be a random variable such that $EX < 0$ and such that for every $\epsilon > 0$ there is $c > 0$ such that for all x large enough,*

$$P(X > x) \leq c x^{-\beta+\epsilon} \tag{2.2}$$

for $\beta > 1$. Suppose that for each $\gamma > 0$, $X^{(\gamma)}$ is a random variable such that $X^{(\gamma)} \stackrel{\mathrm{st}}{\leq} X$, and such that for some $d > 3$ and $h > 0$

$$P\Big(X^{(\gamma)} > \frac{1}{d}\gamma\Big) \leq c\gamma^{-h} \tag{2.3}$$

for all γ large enough.

Let $a \geq 0$. Define a family of Markov chains by $Z_1^{(\gamma)} = a$, and

$$Z_{n+1}^{(\gamma)} = \max(Z_n^{(\gamma)} + X_n^{(\gamma)}, a), \quad n \geq 1,$$

where $(X_n^{(\gamma)}, n \geq 1)$ are i.i.d. copies of $X^{(\gamma)}$. Let

$$n_\gamma = \inf\{n \geq 1 : Z_n^{(\gamma)} \geq \gamma\}.$$

Then for every $\delta > 0$ and $0 < \rho < \beta - 1$

$$\varliminf_{\gamma\to\infty} P\Big(n_\gamma > \delta\gamma^{\min(h/2, \rho d/3)}\Big) > 0. \tag{2.4}$$

Proof. We call a cycle the interval of time between successive returns of $Z_n^{(\gamma)}$ to a. Denote the initial cycle length by C_1. Since the length of each cycle is at least 1, for every $\gamma > a$ we have

$$n_\gamma \stackrel{\mathrm{st}}{\geq} G_\gamma, \tag{2.5}$$

where G_γ is a geometric random variable with probability for success given by

$$p_\gamma = P[\bigvee_{j=1}^{C_1} Z_n^{(\gamma)} \geq \gamma], \tag{2.6}$$

the probability that $\{Z_n^{(\gamma)}\}$ reaches or exceeds γ within a cycle. So G_γ is the number of failures before the first success where success means a cycle maximum exceeds γ.

We now estimate p_γ. We have

$$\begin{aligned}
p_\gamma &= P[\bigvee_{n=1}^{C_1} Z_n^{(\gamma)} \geq \gamma] \\
&\leq P[\bigvee_{n=1}^{C_1} Z_n^{(\gamma)} \geq \gamma, C_1 > \gamma^{h/2}] + P[\bigvee_{n=1}^{C_1} Z_n^{(\gamma)} \geq \gamma, C_1 \leq \gamma^{h/2}] \\
&\leq P[C_1 > \gamma^{h/2}] + P[\bigvee_{n=1}^{C_1} Z_n^{(\gamma)} \geq \gamma, C_1 \leq \gamma^{h/2}, \bigvee_{j=1}^{C_1} X_j^{(\gamma)} \geq \frac{\gamma}{d}] \\
&\quad + P[\bigvee_{n=1}^{C_1} Z_n^{(\gamma)} \geq \gamma, C_1 \leq \gamma^{h/2}, \bigvee_{j=1}^{C_1} X_j^{(\gamma)} < \frac{\gamma}{d}] \\
&\leq P[C_1 > \gamma^{h/2}] + P[\bigvee_{j=1}^{[\gamma^{h/2}]} X_j^{(\gamma)} \geq \frac{\gamma}{d}] + P[\bigvee_{n=1}^{C_1} Z_n^{(\gamma)} \geq \gamma, \bigvee_{j=1}^{C_1} X_j^{(\gamma)} < \frac{\gamma}{d}] \\
&=: P_1(\gamma) + P_2(\gamma) + P_3(\gamma).
\end{aligned} \tag{2.7}$$

The assumption $X^{(\gamma)} \overset{\text{st}}{\leq} X$ implies that the cycle length for $(Z_n^{(\gamma)},\, n \geq 1)$ is stochastically dominated by the time of the first return to zero of the Lindley process

$$S_0 = 0, \quad S_{n+1} = (S_n + X_{n+1})_+, \quad n \geq 0, \tag{2.8}$$

where $(X_n,\, n \geq 1)$ are i.i.d. copies of X. Since the latter time of the first return to zero has a finite mean because of the negative mean of X (see e.g. Proposition 7.6.4 of Resnick (1992)) we use Markov's inequality to see that

$$P_1(\gamma) \leq c\gamma^{-h/2}. \tag{2.9}$$

Furthermore, by (2.3) we have

$$P_2(\gamma) \leq \gamma^{h/2} P\Big(X^{(\gamma)} > \frac{1}{d}\gamma\Big) \leq c\gamma^{-h/2}. \tag{2.10}$$

Finally, to estimate $P_3(\gamma)$ let us consider an unconstrained random walk defined by $\tilde{Z}_1^{(\gamma)} = 0$, and

$$\tilde{Z}_{n+1}^{(\gamma)} = \tilde{Z}_n^{(\gamma)} + X_n^{(\gamma)}, \quad n \geq 1.$$

Then

(2.11)
$$P_3(\gamma) \leq P\Big(\sup_{n\geq 1} \tilde{Z}_n^{(\gamma)} \geq \gamma - a, \text{ and the level } \gamma - a \text{ is initially reached without steps } \geq \frac{1}{d}\gamma\Big).$$

Let

$$m_1(\gamma) = \inf\{n \geq 1 : \tilde{Z}_n^{(\gamma)} \geq \frac{1}{d}\gamma\}.$$

For $j \geq 2$ we define $m_j(\gamma) = \infty$ if $m_{j-1}(\gamma) = \infty$, whereas if $m_{j-1}(\gamma) < \infty$, we define

$$m_j(\gamma) = \inf\{n > m_{j-1}(\gamma) : \tilde{Z}_n^{(\gamma)} - \tilde{Z}_{m_{j-1}(\gamma)}^{(\gamma)} \geq \frac{1}{d}\gamma\}.$$

Then it follows from (2.11) that for all γ large enough (comparatively to a),

$$P_3(\gamma) \leq P\Big(m_{[d/3]}(\gamma) < \infty\Big) = \Big(P(m_1(\gamma) < \infty)\Big)^{[d/3]} \tag{2.12}$$

by the strong Markov property of the random walk. To check the first inequality in (2.12), set

$$T(\gamma - a) = \inf\{n : \tilde{Z}_n^{(\gamma)} \geq \gamma - a\}.$$

Observe if the steps $X_n^{(\gamma)}$ are bounded by γ/d, then at time $m_1(\gamma)$, the maximum height of the random walk $\{Z_n^{(\gamma)}\}$ is

$$\frac{\gamma}{d} + \frac{\gamma}{d} = \frac{2\gamma}{d}.$$

Similarly, if $m_j(\gamma) < \infty$, then after j steps, the maximum height of the random walk $\{Z_n^{(\gamma)}\}$ is $j \cdot 2\gamma/d$. Therefore

$$
\begin{aligned}
\big[m_{[d/3]}(\gamma) = \infty, \bigvee_{n=1}^{T(\gamma-a)} X_n^{(\gamma)} \leq \frac{\gamma}{d}\big] \\
&\subset \big[\text{ the maximum height of the random walk is } \left(\left[\frac{d}{3}\right] - 1\right)\frac{2\gamma}{d} \\
&\qquad \text{at time } m_{[d/3]-1}(\gamma) \text{ and the random walk} \\
&\qquad \text{never goes } \gamma/d \text{ higher, } \bigvee_{n=1}^{T(\gamma-a)} X_n^{(\gamma)} \leq \frac{\gamma}{d}\big] \\
&\subset \big[\bigvee_{n=1}^{\infty} Z_n^{(\gamma)} \leq \frac{2\gamma}{3} + \frac{\gamma}{d}, \bigvee_{n=1}^{T(\gamma-a)} X_n^{(\gamma)} \leq \frac{\gamma}{d}\big] \\
&\subset \big[\bigvee_{n=1}^{\infty} Z_n^{(\gamma)} < \gamma - a, \bigvee_{n=1}^{T(\gamma-a)} X_n^{(\gamma)} \leq \frac{\gamma}{d}\big]
\end{aligned}
$$

provided

$$\gamma\left(\frac{2}{3} + \frac{1}{d}\right) < \gamma - a$$

or

$$a < \gamma\left(\frac{1}{3} - \frac{1}{d}\right).$$

Observe that it follows from the assumption $X^{(\gamma)} \overset{\text{st}}{\leq} X$ that

$$P(m_1(\gamma) < \infty) \leq P\Big(\sup_{n \geq 0} \tilde{S}_n > \frac{1}{d}\gamma\Big), \tag{2.13}$$

where $(\tilde{S}_n,\, n \geq 0)$ is the unconstrained version of the random walk in (2.8): $\tilde{S}_0 = 0$ and

$$\tilde{S}_{n+1} = \tilde{S}_n + X_{n+1}, \quad n \geq 0.$$

Fix a $0 < \rho < \beta - 1$. It follows from (2.2) that there is a random variable Y such that $EY < 0$, $Y \overset{\text{st}}{\geq} X$, and

$$P(Y > y) \ \text{ is regularly varying at infinity with exponent } \rho + 1. \tag{2.14}$$

Then

$$P\Big(\sup_{n \geq 0} \tilde{S}_n > \frac{1}{d}\gamma\Big) \leq P\Big(\sup_{n \geq 0} S_n^* > \frac{1}{d}\gamma\Big), \tag{2.15}$$

where $(S_n^*,\, n \geq 0)$ is the walk $S_0^* = 0$ and

$$S_{n+1}^* = S_n^* + Y_{n+1}, \quad n \geq 0,$$

where $(Y_n,\, n \geq 1)$ are i.i.d. copies of Y. Now, it follows from Embrechts and Goldie (1982) that for a negative mean random walk satisfying (2.14) we have

$$P\Big(\sup_{n\geq 0} S_n^* > \frac{1}{d}\gamma\Big) \text{ is regularly varying at infinity with exponent } \rho. \tag{2.16}$$

We conclude by (2.12), (2.13), (2.15) and (2.16) that for every $0<\rho<\beta-1$,

$$P_3(\gamma) \leq c\gamma^{-\rho d/3} \tag{2.17}$$

for all γ large enough. We conclude by (2.7), (2.9), (2.10) and (2.17) that for any $0<\rho<\beta-1$,

$$p_\gamma \leq c\gamma^{-\min(h/2,\rho d/3)} \tag{2.18}$$

for all γ large enough.

In particular, it follows from (2.18) that $p_\gamma \to 0$ as $\gamma \to \infty$. Therefore, $p_\gamma G_\gamma$ converges weakly, as $\gamma \to \infty$, to a mean 1 exponential random variable. We conclude that

$$\begin{aligned}\varliminf_{\gamma\to\infty} P\Big(n_\gamma > \delta\gamma^{\min(h/2,\rho d/3)}\Big) &\geq \varliminf_{\gamma\to\infty} P\Big(G_\gamma > \delta\gamma^{\min(h/2,\rho d/3)}\Big)\\ &= \varliminf_{\gamma\to\infty} P\Big(p_\gamma G_\gamma > \delta p_\gamma\gamma^{\min(h/2,\rho d/3)}\Big) > 0\end{aligned}$$

by (2.18). This completes the proof of the lemma. □

Our next sequence of results deals with certain aspects of positive dependence occurring in an $M/G/\infty$ queue.

Lemma 3. *Let $(Z_i,\, i \geq 1)$ be the sequence of successive session lengths (indexed by the order of their arrivals) in the $M/G/\infty$ model underlying (1.1). For $i = 1, 2, \ldots$ let K_i denote the number of new sessions arriving during the ith session (of length Z_i). Then for every $m \geq 1$ and $n_i = 1, 2, \ldots,\ i = 1, \ldots, m$ we have*

$$P(K_1 < n_1, \ldots, K_m < n_m) \geq \prod_{i=1}^m P(K_i < n_i). \tag{2.19}$$

Proof. Let the Poisson arrival stream of sessions be $\{T_i, i \geq 1\}$ where $0 < T_1 < T_2 < \ldots$ so that

$$N(A) = \sum_{i=1}^\infty \epsilon_{T_i}(A)$$

is the number of sessions initiated in a set A. The point process N is *associated* (Burton and Waymire, 1985, page 1271), (Resnick, 1987, page 300) and so for intervals $I_1, \ldots, I_m$ in $(0,\infty)$ and nonnegative integers $l_1, l_2, \ldots, l_m$

$$\begin{aligned}&P\Big(N(I_i) < l_i, i = 1, \ldots, m\Big)\\ &\geq P\Big(N(I_i) < l_i,\, i = 1, \ldots, m_1\Big) P\Big(N(I_i) < l_i,\, i = m_1+1, \ldots, m\Big).\end{aligned} \tag{2.20}$$

Another reference for other facts on associated random variables is Esary et al. (1967).

The proof of (2.19) is by induction on m. Since there is nothing to prove for $m = 1$, assume that (2.19) holds for some $m \geq 1$, and let us prove it for $m+1$. For fixed $n_i \geq 1$, $i = 1, \ldots, m$

consider a set in $\mathbb{R}_+^{2m+1}$ defined by

$$B_{m,n_1,\dots,n_m} = \Big\{(t_1,\dots,t_m,t_{m+1},z_1,\dots,z_m) :\ 0<t_1<\dots<t_m<t_{m+1},\ z_1\ge 0,\dots,z_m\ge 0,\\ \text{and after the first } m+1 \text{ arrivals of sessions at times } t_1,\dots,t_m,t_{m+1},\\ \text{the lengths of the first } m \text{ of which are } z_1,\dots,z_m, \text{ the condition}\\ K_1<n_1,\dots,K_m<n_m \text{ is not yet violated}\Big\}.$$

We have then

$$\begin{aligned} &P(K_1<n_1,\dots,K_m<n_m,K_{m+1}<n_{m+1})\\ &= \int_{B_{m,n_1,\dots,n_m}} \lambda^{m+1}e^{-\lambda t_{m+1}}\,dt_1\dots dt_m\,dt_{m+1}\,F(dz_1)\dots F(dz_m)\\ &\quad\times\int_0^\infty P\Big(\bigcap_{l=1}^{m+1}[K_l<n_l]\Big|T_i=t_i,\ Z_i=z_i,\ i=1,\dots,m+1\Big)\,F(dz_{m+1}), \end{aligned} \tag{2.21}$$

where T_i stands for the arrival time of the ith session. Now, for fixed $t_1,\dots,t_m,t_{m+1},z_1,\dots,z_m\in B_{m,n_1,\dots,n_m}$ let $I_i=(t_{m+1},t_i+z_i)$ ($=\emptyset$ if $t_{m+1}\ge t_i+z_i$), $i=1,\dots,m$. By the definition of $B_{m,n_1,\dots,n_m}$ there are $l_1,\dots,l_m$ such that

$$\begin{aligned} &P\Big(K_1<n_1,\dots,K_m<n_m,K_{m+1}<n_{m+1}\Big|\,T_i=t_i,\ Z_i=z_i,\ i=1,\dots,m+1\Big)\\ &= P\Big(N(I_i)<l_i,\ i=1,\dots,m,\ N\big((t_{m+1},t_{m+1}+z_{m+1})\big)<n_{m+1}\Big)\\ &\ge P\Big(N\big((t_{m+1},t_{m+1}+z_{m+1})\big)<n_{m+1}\Big)P\Big(N(I_i)<l_i,\ i=1,\dots,m\Big)\\ &= P\Big(N\big((t_{m+1},t_{m+1}+z_{m+1})\big)<n_{m+1}\Big)\\ &\quad\times P\Big(K_1<n_1,\dots,K_m<n_m\Big|\,T_i=t_i,\ Z_i=z_i,\ i=1,\dots,m+1\Big), \end{aligned}$$

where we have used (2.20). Substituting the above in (2.21) we obtain

$$\begin{aligned} &P(K_1<n_1,\dots,K_m<n_m,K_{m+1}<n_{m+1})\\ &\ge \int_{B_{m,n_1,\dots,n_m}} P\Big(K_1<n_1,\dots,K_m<n_m\Big|\,T_i=t_i,\ Z_i=z_i,\ i=1,\dots,m+1\Big)\\ &\quad\times\lambda^{m+1}e^{-\lambda t_{m+1}}\,dt_1\dots dt_m\,dt_{m+1}\,F(dz_1)\dots F(dz_m)\\ &\quad\int_0^\infty P\Big(N\big((t_{m+1},t_{m+1}+z_{m+1})\big)<n_{m+1}\Big)\,F(dz_{m+1})\\ &= \int_{B_{m,n_1,\dots,n_m}} P\Big(K_1<n_1,\dots,K_m<n_m\Big|\,T_i=t_i,\ Z_i=z_i,\ i=1,\dots,m+1\Big)\\ &\quad\lambda^{m+1}e^{-\lambda t_{m+1}}\,dt_1\dots dt_m\,dt_{m+1}\,F(dz_1)\dots F(dz_m)P(K_{m+1}<n_{m+1})\\ &= P(K_1<n_1,\dots,K_m<n_m)P(K_{m+1}<n_{m+1})\\ &\ge \prod_{i=1}^{m+1} P(K_i<n_i) \end{aligned}$$

by the assumption of the induction. This completes the proof. ∎

Lemma 4. *Fix an $m \geq 1$, and let*

$$E_i = \Big\{ \textit{during the ith session, the number of simultaneously present newly arriving sessions is always less than } m \Big\}, \quad i = 1, 2, \dots . \tag{2.22}$$

Then for every $n \geq 1$,

$$P(E_1 \cap \cdots \cap E_n) \geq \Big(P(E_1)\Big)^n. \tag{2.23}$$

Proof. We start, once again, with a simple association statement. The process $\{N(t),\, t \geq 0\}$ describing the number of customers in the system in the $M/G/\infty$ queue consists of associated random variables. To see this, endow $M_p([0,\infty) \times [0,\infty])$, the space of point measures on $[0,\infty) \times [0,\infty]$, with the partial order "$\leq$" defined by

$$\nu \leq \mu$$

iff

$$\nu(A) \leq \mu(A), \quad \forall A \in \mathcal{B}([0,\infty) \times [0,\infty]),$$

which means the support of ν is contained in the support of μ. If

$$\psi_t : (M_p([0,\infty) \times [0,\infty]), \leq) \mapsto (\mathbb{R}, \leq)$$

is monotone, then since

$$M := \sum_{k=1}^{\infty} \epsilon_{(T_k, Z_k)}$$

is a Poisson process on $[0,\infty) \times [0,\infty]$, M is associated (Burton and Waymire (1985), Resnick (1987)) and thus $(\psi_t(M), t > 0)$ is an associated family. It remains to observe that if $\nu \in M_p([0,\infty) \times [0,\infty])$, the map

$$\psi_t(\nu) = \nu\{(s,z) : s \leq t \leq s+z\}$$

is monotone and

$$\psi_t(M) = N(t).$$

So we conclude $(N(t), t > 0)$ is an associated family of random variables.

We now prove (2.23), and the proof is by induction on n. For $n = 1$ there is nothing to prove, so assume that (2.23) holds for some $n \geq 1$, and let us prove it for $n+1$. In the following computation $t_2, \dots, t_{n+1}$ stand for realizations of the arrival times of the next n sessions after the start of the first session (which is assumed to have arrived at time 0), and $z_1, z_2, \dots, z_{n+1}$ stand for the corresponding lengths of the first $n+1$ sessions. Recall that F is the distribution of the session lengths. We have

$$P(E_1 \cap \cdots \cap E_n \cap E_{n+1}) = \int_{[0,\infty)^n} F(dz_1) \dots F(dz_n) \int_{0 < t_2 < \dots < t_{n+1}} \lambda^n e^{-\lambda t_{n+1}}\, dt_2 \dots dt_{n+1} \int_0^\infty F(dz_{n+1}) P\Big(A_1(z_1, \dots, z_n, t_2, \dots, t_{n+1}; z_{n+1}) \cap A_2(z_{n+1})\Big), \tag{2.24}$$

where $A_1(z_1, \dots, z_n, t_2, \dots, t_{n+1}; z_{n+1})$ is the event that the number of simultaneously present newly arriving sessions during each one of the first n sessions is always less than m and when arrivals $2, \dots, n+1$ occur at $t_2, \dots, t_{n+1}$ and lengths of the first $n+1$ sessions are $z_1, \dots, z_{n+1}$.

Also define $A_2(z_{n+1})$ to be the event that the number of simultaneously present newly arriving sessions during the time z_{n+1} is always less than m. The event $A_1(z_1,\dots,z_n,t_2,\dots,T_{n+1};z_{n+1})$ is, of course, determined by the sessions arriving after time t_{n+1}, but for some combination of the parameters $z_1,\dots,z_n,t_2,\dots,t_{n+1}$ and z_{n+1} the conditions describing the event may already be violated by the time t_{n+1}, so no choice of new arrivals can, in that case, make the event happen.

For $t\ge 0$, let $N_1(t)$ denote the number of sessions that arrived to the system after the time t_{n+1} and that are active at time $t+t_{n+1}$. The process $\{N_1(t),\, t\ge 0\}$ has the same law as the process $\{N(t),\, t\ge 0\}$, and, hence, is associated. Furthermore, it is obvious that the indicator functions of the events $A_1(z_1,\dots,z_n,t_2,\dots,t_{n+1};z_{n+1})$ and $A_2(z_{n+1})$ are non-increasing functions of each of $N_1(t)$'s. Therefore, these random variables are associated, and so

$$
\begin{aligned}
(2.25)\qquad & P\Big(A_1(z_1,\dots,z_n,t_2,\dots,t_{n+1};z_{n+1})\cap A_2(z_{n+1})\Big)\\
&\qquad \ge P\Big(A_1(z_1,\dots,z_n,t_2,\dots,t_{n+1};z_{n+1})\Big)P\Big(A_2(z_{n+1})\Big).
\end{aligned}
$$

Observe that the probability $P\Big(A_1(z_1,\dots,z_n,t_2,\dots,t_{n+1};z_{n+1})\Big)$ is, for fixed $z_1,\dots,z_n$ and $t_2,\dots,t_{n+1}$, a non-increasing function of z_{n+1}. Furthermore, the probability $P\Big(A_2(z_{n+1})\Big)$ is a non-increasing function of z_{n+1} as well. Therefore, for fixed $z_1,\dots,z_n,t_2,\dots,t_{n+1}$, if we consider $P(A_1(z_1,\dots,z_n,t_2,\dots,t_{n+1};\cdot))$ and $P(A_2(\cdot))$ as random variables on $(\mathbb{R}_+,\mathcal{B},F)$, they will be associated. We conclude from this and from (2.24) and (2.25) that

$$
\begin{aligned}
&P(E_1\cap\cdots\cap E_n\cap E_{n+1})\\
&\quad\ge \int_{[0,\infty)^n} F(dz_1)\dots F(dz_n)\int_{0<t_2<\dots<t_{n+1}} \lambda^n e^{-\lambda t_{n+1}}\,dt_2\dots dt_{n+1}\\
&\qquad \int_0^\infty F(dz_{n+1})P\Big(A_1(z_1,\dots,z_n,t_2,\dots,t_{n+1};z_{n+1})\Big)P\Big(A_2(z_{n+1})\Big)\\
&\quad\ge \int_{[0,\infty)^n} F(dz_1)\dots F(dz_n)\int_{0<t_2<\dots<t_{n+1}} \lambda^n e^{-\lambda t_{n+1}}\,dt_2\dots dt_{n+1}\\
&\qquad \int_0^\infty P\Big(A_1(z_1,\dots,z_n,t_2,\dots,t_{n+1};z_{n+1})\Big)F(dz_{n+1})\int_0^\infty P\Big(A_2(z_{n+1})\Big)F(dz_{n+1})\\
&\quad= \int_{[0,\infty)^n} F(dz_1)\dots F(dz_n)\int_{0<t_2<\dots<t_{n+1}} \lambda^n e^{-\lambda t_{n+1}}\,dt_2\dots dt_{n+1}\\
&\qquad \int_0^\infty P\Big(A_1(z_1,\dots,z_n,t_2,\dots,t_{n+1};z_{n+1})\Big)F(dz_{n+1})P(E_1)\\
&\quad= P(E_1\cap\dots\cap E_n)P(E_1)\\
&\quad\ge \Big(P(E_1)\Big)^{n+1}
\end{aligned}
$$

by the assumption of the induction. This completes the proof. □

If $(\Omega,\mathcal{F},P)$ is a probability space, $A_i\in\mathcal{F}$ for $i\ge 1$, and N a random variable with nonnegative integer values on this probability space, then we use the notation

$$\cup_{i=1}^N A_i := \cup_{i=1}^\infty\Big(A_i\cap\{N\ge i\}\Big).$$

In the following lemma the events E_i are defined by (2.22).

Lemma 5. *For every $t > 0$*

$$P\Big(\cup_{i=1}^{N(t)} E_i^c\Big) \le P(E_1^c)E(N(t)). \tag{2.26}$$

Proof. Our claim (2.26) follows from the obvious fact that

$$P\Big(\cup_{i=1}^{N(t)} E_i^c\Big) \le E\Big(\sum_{i=1}^{N(t)} \mathbf{1}(E_i^c)\Big)$$

and the following version of Wald's identity.

Let $(W_i,\, i \ge 1)$ be identically distributed random variables with a finite mean on a probability space $(\Omega, \mathcal{F}, P)$, and N a random variable with nonnegative integer values on the same probability space. Assume that

$$E(W_i|N=n) = E(W_i) \quad \text{for all } n < i. \tag{2.27}$$

Then

$$E(\sum_{i=1}^{N} W_i) = E(W_1)\, E(N). \tag{2.28}$$

Indeed,

$$\begin{aligned}
E(\sum_{i=1}^{N} W_i) &= \sum_{n=0}^{\infty} P(N=n)E(\sum_{i=1}^{n} W_i|N=n) \\
&= \sum_{n=1}^{\infty} P(N=n) \sum_{i=1}^{n} E(W_i|N=n) \\
&= \sum_{i=1}^{\infty} \sum_{n=i}^{\infty} P(N=n)E(W_i|N=n) \\
&= \sum_{i=1}^{\infty} \Big(\sum_{n=0}^{\infty} P(N=n)E(W_i|N=n) - \sum_{n=0}^{i-1} P(N=n)E(W_i|N=n)\Big) \\
&= \sum_{i=1}^{\infty} \Big(E(W_i) - E(W_i) \sum_{n=0}^{i-1} P(N=n)\Big) \\
&= EW_1 \sum_{i=1}^{\infty} P(N \ge i) = E(W_1)\, EN,
\end{aligned}$$

proving (2.28). □

Our next lemma gives a bound on the order of magnitude of the probability of the event E_1 in (2.22) for a choice of parameters we will need in the sequel.

Lemma 6. *Let F be a distribution whose tail $1-F$ is dominatedly varying with Matuszewska index $\beta > 1$ in (1.6). For $\gamma > 0$ we let the distribution of the session lengths F_γ be defined by*

$$\overline{F_\gamma}(x) = \frac{\overline{F}(x)}{\overline{F}(\gamma)},\ x \ge \gamma.$$

Assume, further, that the intensity of the Poisson process of arriving sessions is $\lambda\overline{F}(\gamma)$. Then there is a finite positive constant $c = c(m)$ (m is the parameter defining E_i in (2.22)) such that for all $\gamma \geq 1$

$$P_\gamma(E_1) \geq 1 - c\Big(\gamma\overline{F}(\gamma)\Big)^m. \tag{2.29}$$

Here P_γ means that the corresponding probability is computed for a system defined using this particular γ.

Proof. The proof is by induction in m. Take first $m = 1$. We have

$$\begin{aligned}1 - P_\gamma(E_1) &= \frac{1}{\overline{F}(\gamma)}\int_\gamma^\infty \Big(1 - e^{-\lambda z\overline{F}(\gamma)}\Big)\, F(dz) \\ &= \lambda\overline{F}(\gamma)\int_0^\gamma e^{-\lambda\overline{F}(\gamma)x}\, dx + \lambda\int_\gamma^\infty \overline{F}(x)e^{-\lambda\overline{F}(\gamma)x}\, dx \\ &=: I_1(\gamma) + I_2(\gamma).\end{aligned}$$

Clearly,

$$I_1(\gamma) \leq \lambda\gamma\overline{F}(\gamma).$$

Furthermore, by (1.6) we have for all $\gamma \geq x_0$ and any $0 < \epsilon < \beta - 1$

$$\begin{aligned}I_2(\gamma) &= \lambda\gamma\int_1^\infty \overline{F}(\gamma y)e^{-\lambda\overline{F}(\gamma)\gamma y}\, dy \\ &\leq C\lambda\gamma\overline{F}(\gamma)\int_1^\infty y^{-(\beta-\epsilon)}\, dy \\ &= c\gamma\overline{F}(\gamma)\end{aligned}$$

for some $0 < c < \infty$. Therefore, we have (2.29) for $m = 1$.

Assume now that (2.29) holds for an $m \geq 1$, and let us prove it for $m + 1$. We have

$$\begin{aligned}1 - P_\gamma(E_1) = \frac{1}{\overline{F}(\gamma)}\int_\gamma^\infty F(dz)\, P_\gamma\Big(&\text{within time interval } (0,z) \text{ there is a time when} \\ &\text{at least } m+1 \text{ newly arrived sessions are simultaneously active}\Big) \\ \leq \frac{1}{\overline{F}(\gamma)}\int_\gamma^\infty F(dz)\, P_\gamma\Big(&\text{within at least one of } N(0,z] \text{ new sessions arriving in } (0,z), \\ &\text{the number of simultaneously present subsequently arriving sessions is at least } m\Big).\end{aligned} \tag{2.30}$$

However, by Lemma 5 and the assumption of the induction we have

$$\begin{aligned}P\Big(&\text{within at least one of } N(0,z], \text{ new sessions arriving in } (0,z) \text{ the number} \\ &\text{of simultaneously present subsequently arriving sessions is at least } m\Big) \\ &\leq EN(0,z]P_\gamma\Big(\text{during a session, the number of simultaneously present} \\ &\qquad \text{subsequently arriving sessions is at least } m\Big) \\ &\leq c\overline{F}(\gamma)z\Big(\gamma\overline{F}(\gamma)\Big)^m.\end{aligned}$$

Substituting the above bound into (2.30), we obtain

$$1 - P_\gamma(E_1) \le c\Big(\gamma\overline{F}(\gamma)\Big)^m \int_\gamma^\infty z\, F(dz). \tag{2.31}$$

However, the same easy computation we used above in the case $m = 1$ shows that

$$\int_\gamma^\infty z\, F(dz) \le c\gamma\overline{F}(\gamma),$$

all γ big enough. Substituting the above into (2.31) completes the inductive step. □

Lemma 7. *Let $X_1, \dots, X_n$ be independent random variables, and let A be a measurable increasing set in R^d (that is, $(x_1, \dots, x_n) \in A$ and $y_i \ge x_i$, $i = 1, \dots, n$ implies $(y_1, \dots, y_n) \in A$.) Then for any u*

$$P\Big(\sum_{i=1}^n X_i > u,\ (X_1, \dots, X_n) \in A\Big) \ge P\Big(\sum_{i=1}^n X_i > u\Big)P\Big((X_1, \dots, X_n) \in A\Big).$$

Proof. The random variables $X_1, \dots, X_n$ are independent, hence associated. If $V_i : \mathbb{R}^n \mapsto \mathbb{R}$, $i = 1, 2$ are nonincreasing functions, then $V_1(X_1, \dots, X_n), V_2(X_1, \dots, X_n)$ are associated. Choose

$$V_1(x_1, \dots, x_n) = \mathbf{1}(\sum_{i=1}^n x_i > u), \quad V_2(x_1, \dots, x_n) = \mathbf{1}((x_1, \dots, x_n) \in A)$$

and check both are monotone. The statement of the lemma now follows. □

3. Bounds on the expected hitting time

The following is the main result of this paper.

Theorem 1. *Assume that the session length distribution F has a dominatedly varying tail, with Matuszewska indices in (1.6) satisfying $\beta > 1$. Assume that (1.2) holds, and let k be defined by (1.4). Assume, further, that $r - \lambda\mu$ is not an integer. Then there is a $C \ge 1$ and a $\gamma_0 > 0$ such that*

$$C^{-1}\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k} \le E\tau_\gamma \le C\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k} \tag{3.1}$$

for all $\gamma \ge \gamma_0$.

It is interesting that Theorem 1 exhibits a "phase transition"–type of dependence for system performance on the service rate r. If one increases r without changing the parameter k, then the asymptotic growth rate of the expected time until overflow, $E\tau_\gamma$, does not change (even though the multiplicative constant may be affected). Hence, the system performance sees little improvement. On the other hand, once the service rate has increased enough to change the parameter k, the rate of growth of $E\tau_\gamma$ increases, and so the system performance sees marked improvement. Indeed, if, for example, $\overline{F}(x) \asymp \mathrm{const}\, x^{-\alpha}$ as $x \to \infty$, $\alpha > 1$, then Theorem 1 says that

$$E\tau_\gamma \asymp C\gamma^{k(\alpha-1)+1}$$

for large γ and some constant C. The effect of k on the system performance is clearly visible.

Note that the assumption $r - \lambda\mu$ not being being an integer is the same as saying that $r - \lambda\mu - (k-1)$ is strictly positive.

While the case $r - \lambda\mu$ being integer is not likely to be of practical importance, the theoretical behavior of $E\tau_\gamma$ in that case is, most likely, affected by additional distributional properties of the session lengths. Note also that the upper bound in (3.1) holds without the assumption that $r - \lambda\mu$ is not an integer.

Unfortunately, the multiplicative constants in (3.1) are difficult to keep track of in the generality of the situation we are considering in Theorem 1. With stronger assumptions on the tail of the session length distribution F (like regular variation) and careful bookkeeping, one should be able to get certain bounds on these constants. In the simplest case, that of $k = 1$, it is proved in Resnick and Samorodnitsky (1997a) that the limit

$$\lim_{\gamma\to\infty} \overline{F}\Big(\frac{\gamma}{1+\lambda\mu - r}\Big) E\tau\gamma$$

exists, and the limit is identified under somewhat more restrictive assumptions on F. We suspect that under appropriate assumptions on F, a similar result should hold for a general k, but the argument has, so far, escaped us.

Proof of Theorem 1. Choose any $\delta > 0$ and take any $n \geq 1$. Fix also $\epsilon_1 > 0$ and $0 < \epsilon_2 < 1$ such that

$$(r-k)\epsilon_1 + \lambda\mu\epsilon_2 \leq \frac{k+\lambda\mu-r}{2}. \tag{3.2}$$

Notice that if for some $i = 0, 1, \dots, n-1$ at least k sessions of length at least $\delta + 2\gamma(1+\epsilon_1)/(k + \lambda\mu - r)$ each arrive in the interval $\big(i\delta, (i+1)\delta\big)$, and if the total length of the sessions of length at most $2\gamma\epsilon_1/(k+\lambda\mu-r)$ arriving in the interval $\big((i+1)\delta, (i+1)\delta + 2\gamma/(k+\lambda\mu-r)\big)$ is at least $2\gamma(1-\epsilon_2)\lambda\mu/(k+\lambda\mu-r)$, then by the time $(i+1)\delta + 2\gamma(1+\epsilon_1)/(k+\lambda\mu-r)$ the amount of work in the buffer is at least

$$\frac{(k-r)2\gamma(1+\epsilon_1)}{k+\lambda\mu-r} + \frac{2\gamma(1-\epsilon_2)\lambda\mu}{k+\lambda\mu-r} = 2\gamma\Big((1+\epsilon_1) - \frac{\lambda\mu}{k+\lambda\mu-r}(\epsilon_1+\epsilon_2)\Big) \geq \gamma$$

where the last inequality uses (3.2). Therefore,

$$\tau_\gamma \leq (i+1)\delta + \frac{2\gamma(1+\epsilon_1)}{k+\lambda\mu-r} \leq n\delta + \frac{2\gamma(1+\epsilon_1)}{k+\lambda\mu-r}.$$

If we choose

$$\delta = M\gamma, \quad \text{with } M > 2/(k+\lambda\mu-r),$$

then the intervals $\big((i+1)\delta, (i+1)\delta + 2\gamma/(k+\lambda\mu-r)\big)$, $i = 0, 1, \dots, n-1$ are disjoint. Denote by A_{ik} the event that at least k sessions of length at least $\delta + 2\gamma(1+\epsilon_1)/(k+\lambda\mu-r)$ arrive in the interval $(i\delta, (i+1)\delta)$. Let B_{ik} be the event that the total length of the sessions of length at most $2\gamma\epsilon_1/(k+\lambda\mu-r)$ arriving in $((i+1)\delta, ((i+1)\delta+2\gamma)/(k+\lambda\mu-r))$ is at least $2\gamma(1-\epsilon_2)\lambda\mu/(k+\lambda\mu-r)$. The previous discussion can be summarized as

$$\bigcup_{i=0}^{n-1}(A_{ik}B_{ik}) \subset \Big[\tau_\gamma \leq n\delta + \frac{2\gamma(1+\epsilon_1)}{k+\lambda\mu-r)}\Big]$$

where $\{(A_{ik}, B_{ik}), 0 \le i \le n-1\}$ are independent. Therefore

$$\begin{aligned} P[\tau_\gamma > n\delta + \frac{2\gamma(1+\epsilon_1)}{k+\lambda\mu - r}] &\le P\left(\bigcap_{i=0}^{n-1}(A_{ik}^c \cup B_{ik}^c)\right) \\ &= (P(A_{ik}^c) + P(B_{ik}^c))^n \\ &= p_\gamma^n := (p_\gamma^{(1)} + p_\gamma^{(2)})^n. \end{aligned}$$

We conclude that

$$\begin{aligned} E\tau_\gamma &= \int_0^\infty P(\tau_\gamma > t)\,dt \\ &\le \Big(\delta + \frac{2\gamma(1+\epsilon_1)}{k+\lambda\mu-r}\Big) + \sum_{n=1}^\infty \int_{n\delta+2\gamma(1+\epsilon_1)/(k+\lambda\mu-r)}^{(n+1)\delta+2\gamma(1+\epsilon_1)/(k+\lambda\mu-r)} P(\tau_\gamma > t)\,dt \\ &\le \Big(\delta + \frac{2\gamma(1+\epsilon_1)}{k+\lambda\mu-r}\Big) + \delta\sum_{n=1}^\infty P\Big(\tau_\gamma > n\delta + 2\gamma(1+\epsilon_1)/(k+\lambda\mu-r)\Big) \qquad (3.3) \\ &\le \Big(\delta + \frac{2\gamma(1+\epsilon_1)}{k+\lambda\mu-r}\Big) + \delta\frac{1}{1-p_\gamma}. \end{aligned}$$

We now estimate $1-p_\gamma$. By the finiteness of the mean of the session length distribution F and the fact that $\delta = M\gamma$ we have

$$\delta\overline{F}\Big(\delta + \frac{2\gamma(1+\epsilon_1)}{k+\lambda\mu-r}\Big) \to 0$$

as $\gamma \to \infty$. Therefore,

$$\begin{aligned} 1 - p_\gamma^{(1)} &= \sum_{j=k}^\infty e^{-\delta\lambda\overline{F}(\delta+2\gamma(1+\epsilon_1)/(k+\lambda\mu-r))}\frac{\Big(\delta\lambda\overline{F}(\delta+\frac{2\gamma(1+\epsilon_1)}{k+\lambda\mu-r})\Big)^j}{j!} \qquad (3.4) \\ &\sim \frac{\Big(\delta\lambda\overline{F}(\delta+\frac{2\gamma(1+\epsilon_1)}{k+\lambda\mu-r})\Big)^k}{k!} \ge c(\gamma\overline{F}(\gamma))^k \end{aligned}$$

for large γ, where we have used (1.6). Recall that here, and in the sequel, c stands for a finite positive constant, whose exact value is not being kept track of, and which may change from time to time. By Lemma 1 we know that, as $\gamma \to \infty$,

$$p_\gamma^{(2)} = o(e^{-c\gamma}), \qquad (3.5)$$

$c > 0$. Therefore, we conclude by (3.3), (3.4) and (3.5) that

$$\varlimsup_{\gamma\to\infty} \gamma^{-1}\Big(\gamma\overline{F}(\gamma)\Big)^k E\tau_\gamma < \infty. \qquad (3.6)$$

This proves the upper bound in (3.1). The lower bound is a bit trickier.

We modify our fluid queue as follows. Instead of a single server with service rate r, we consider two servers, one with service rate $k-1$, and the other with service rate $r_1 = r-(k-1) > 0$, each with an infinite buffer. (Note that since $\lambda\mu + (k-1) \le r$ by definition of k, we have $r_1 > 0$.) We route the arriving sessions as follows. Fix some $K > 0$ (to be specified later). All the work arriving in the sessions whose length exceeds γ/K goes into the buffer of server 1, while all the rest goes into the buffer of server 2. The state of the new system is, by definition, the combined amount of work in the two buffers. It is clear that the new system is less efficient than the original one, and so the state of the new system will not reach level γ at a later time than the original system. We will use the same notation, τ_γ, to describe the first time the state of the new system reaches γ. Since we will work only with the new system until the end of the proof of the theorem (unless stated otherwise), no confusion should result from this ambiguity in notation. To prove the lower bound in (3.1) we need to prove that (for the new system)

$$\varliminf_{\gamma\to\infty} \gamma^{-1}\Big(\gamma\overline{F}(\gamma)\Big)^{k} E\tau_\gamma > 0. \tag{3.7}$$

Let $X_i(t)$ denote the content of the buffer of server $i = 1, 2$ at time $t \geq 0$, and let

$$\tau_\gamma^{(1)} = \inf\{t \geq 0 :\ X_1(t) > 0\}$$

and

$$\tau_\gamma^{(2)} = \inf\{t \geq 0 :\ X_2(t) \geq \gamma\}.$$

Thus $X_1(\cdot), X_2(\cdot)$ are independent and hence so are $\tau_\gamma^{(1)}$ and $\tau_\gamma^{(2)}$. Then $\tau_\gamma \geq \min(\tau_\gamma^{(1)}, \tau_\gamma^{(2)})$, and so for every $\epsilon > 0$

$$\begin{aligned} E\tau_\gamma &\geq E\min(\tau_\gamma^{(1)}, \tau_\gamma^{(2)}) \geq \epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k} P\Big(\min(\tau_\gamma^{(1)}, \tau_\gamma^{(2)}) \geq \epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big) \\ &= \epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k} P\Big(\tau_\gamma^{(1)} \geq \epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big) P\Big(\tau_\gamma^{(2)} \geq \epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big). \end{aligned} \tag{3.8}$$

It follows from (3.7) and (3.8) that to complete the proof of the theorem it is enough to show that there is an $\epsilon > 0$ and a $\delta > 0$ such that for all γ large enough,

$$P\Big(\tau_\gamma^{(1)} \geq \epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big) \geq \delta \tag{3.9}$$

and

$$P\Big(\tau_\gamma^{(2)} \geq \epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big) \geq \delta. \tag{3.10}$$

We, actually, prove (3.10) first. That is, we concentrate on a system consisting of single server with service rate $r_1 = r - (k-1)$, which serves all sessions whose length does not exceed γ/K. That is, now the term "system" corresponds to this particular system describing the environment of server 2 and "sessions" are only sessions of length not exceeding γ/K. Let us denote by $(U_n,\ n \geq 1)$ the increasing sequence of times the number of active sessions in this system changes from 1 to 0. These times are the ends of busy periods in the underlying $M/G/\infty$ queue. We also refer to these times as ends of activity periods.

If $S_n = X_2(U_n)$ is the state of the system at time U_n, then $(S_n,\ n \geq 0)$ is a Markov chain with $S_0 = 0$. We also denote

$$M_n = \sup_{U_{n-1}\leq t<U_n} X_2(t), \quad n \geq 1,$$

with $U_0 = 0$. M_n is the maximal level the amount of work in the system reaches during the nth activity period of the underlying $M/G/\infty$ queue. Letting

$$n_\gamma = \inf\{n \geq 1 :\ M_n \geq \gamma\}$$

we see that

$$\tau_\gamma^{(2)} \geq U_{n_\gamma - 1}. \tag{3.11}$$

If we denote by I_n and B_n the lengths of the nth idle and busy periods of the underlying $M/G/\infty$ queue (that is, $U_n - U_{n-1} = I_n + B_n$), then it follows from (3.11) that for all $u > 0$ and $m \geq 1$

$$P(\tau_\gamma^{(2)} \geq u) \geq P\Big(U_{n_\gamma - 1} \geq u\Big) \geq P\Big(\sum_{i=1}^{n_\gamma - 1} I_i > u\Big) \tag{3.12}$$

$$\geq P\Big(n_\gamma > m, \sum_{i=1}^{m} I_i > u\Big).$$

We may, obviously, assume that the idle times $I_1, I_2, \ldots$ are defined on a probability space $(\Omega_1, \mathcal{F}_1, P_1)$, while all the rest of random variables generating the underlying $M/G/\infty$ queue live on another probability space $(\Omega_2, \mathcal{F}_2, P_2)$, and the overall probability space is $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, P_1 \times P_2)$. Observe that for a fixed $\omega_2 \in \Omega_2$ the event $\{n_\gamma > m\}$ depends only on $I_1, \ldots, I_m$, and the indicator of this event is a non-decreasing function of $I_1, \ldots, I_m$. Therefore by Lemma 7

$$P\Big(n_\gamma > m, \sum_{i=1}^{m} I_i > u\Big) = E_2\bigg(P_1\Big(n_\gamma > m, \sum_{i=1}^{m} I_i > u\Big)\bigg) \tag{3.13}$$

$$\geq E_2\bigg(P_1\Big(n_\gamma > m\Big)P_1\Big(\sum_{i=1}^{m} I_i > u\Big)\bigg)$$

$$= E_2\left(P_2(n_\gamma > m)\right) P\left(\sum_{i=1}^{m} I_i > u\right)$$

$$= P\Big(n_\gamma > m\Big)P\Big(\sum_{i=1}^{m} I_i > u\Big). \tag{3.14}$$

It follows from (3.12) and (3.13) that

$$P\Big(\tau_\gamma^{(2)} \geq \epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big) \geq P\Big(n_\gamma > 2\epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big)P\Big(\sum_{i \leq 2\epsilon\gamma\left(\gamma\overline{F}(\gamma)\right)^{-k}} I_i > \epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big).$$

However, by the law of large numbers,

$$P\Big(\sum_{i \leq 2\epsilon\gamma\left(\gamma\overline{F}(\gamma)\right)^{-k}} I_i > \epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big) \to 1$$

as $\gamma \to \infty$. Therefore, (3.10) will follow once we prove that

$$\varliminf_{\gamma \to \infty} P\Big(n_\gamma > 2\epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big) > 0. \tag{3.15}$$

Let us denote by F_n the nth time the number of active sessions in the system changes from 0 to 1. These times are the beginnings of activity periods in the underlying $M/G/\infty$ queue. Let $Z_n = X_2(F_n)$, $n \geq 1$. We agree that the nth activity periods begins at time F_n and ends at time U_n. Let us also denote by W_n the total amount of work brought in the system during the nth activity period, and by V_n the length of that part of the nth activity period B_n when the buffer is not empty. Clearly, $0 \leq V_n \leq B_n$. Denoting $Z_1 = 0$, we see then that $(Z_n,\, n \geq 1)$ is a Markov chain satisfying the recursion

$$Z_{n+1} = (Z_n + W_n - r_1 V_n - r_1 I_n)_+, \quad n \geq 1. \tag{3.16}$$

Let

$$m_\gamma = \inf\{n \geq 1 :\ Z_n \geq \gamma\}.$$

Observe that, if for some k, $Z_k < \gamma/2$ and $M_k \geq \gamma$, then $W_k > \gamma/2$. Therefore,

$$\begin{aligned} P\Big(n_\gamma > 2\epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big) &\geq P\Big(m_{\gamma/2} > 2\epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big) \\ &\quad - P\Big(W_k > \frac{\gamma}{2},\ \text{some } k \leq 2\epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big). \end{aligned} \tag{3.17}$$

Fix an $h > 0$. We now use Proposition 1 of Resnick and Samorodnitsky (1997a), according to which we can and do choose K (used in the routing the sessions arriving in the system) so big that

$$P(W_k > u) = o(u^{-h}) \ \ \text{as } u \to \infty. \tag{3.18}$$

If we choose $h > k(\alpha - 1) + 1$, where α is a Matuszewska index in (1.6), then

$$P\Big(W_k > \frac{\gamma}{2},\ \text{some } k \leq 2\epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big) \leq 2\epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k} P\Big(W_1 > \frac{\gamma}{2}\Big) \to 0$$

as $\gamma \to \infty$. Therefore, (3.15) will follow once we prove that for some $\epsilon > 0$

$$\varliminf_{\lambda\to\infty} P\Big(m_\gamma > \epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big) > 0. \tag{3.19}$$

Note that we have used the dominated variation to simplify the above expression.

To analyze the Markov chain (3.16) we note that the random variables $(W_n,\, n \geq 1)$ and $(I_n,\, n \geq 1)$ form two independent sequences of i.i.d. random variables. Unfortunately, the sequence $(V_n,\, n \geq 1)$ is not, in general, an i.i.d. sequence (it is i.i.d. in the case $0 < r_1 < 1$). Furthermore, V_n depends on Z_{n-1}. To overcome this difficulty we use several simple observations. First of all, we see that

$$E(V_n | Z_{n-1} = a) \uparrow EB_n \ \ \text{as } a \uparrow \infty. \tag{3.20}$$

Let us denote by W_n^0 the amount of work brought in the system during the nth activity period of the original system (i.e. the system we discussed before splitting the arriving sessions into two different streams). Similarly, let B_n^0 and I_n^0 be the corresponding busy and idle times. An elementary renewal theorem gives us the relation

$$EW_n^0 = \lambda\mu E(B_n^0 + I_n^0),$$

and so we have

$$E(W_n^0 - r_1 B_n^0 - r_1 I_n^0) = (\lambda\mu - r_1)E(B_n^0 + I_n^0) < 0 \tag{3.21}$$

by the definition of k. It follows from (3.21) and (3.20) that there is an $a \geq 0$ such that for all $\gamma \geq \gamma_0(K)$ we have

$$E(W_n^0 - r_1 I_n) - r_n E(V_n | Z_{n-1} = a) < 0. \tag{3.22}$$

From that point on we fix an a such that (3.22) holds.

We now modify the Markov chain (3.16) by defining $Z_1^* = a$, and

$$Z_{n+1}^* = \max\Big(Z_n^* + W_n - r_1 V_n - r_1 I_n, a\Big), \quad n \geq 1. \tag{3.23}$$

Intuitively, if at time F_n the amount of work in the buffer is less that a, we add the necessary work to increase the amount of work in the system to a. Alternatively, think of putting into the buffer a false bottom at level a. Observe that if both (3.16) and (3.23) are driven by the same random variables, then we have $Z_n^* \geq Z_n$ for all $n \geq 1$.

Let $\big((W_n, V_n'),\, n \geq 1\big)$ be a sequence of i.i.d. random vectors such that the distribution of (W_1, V_1') is the same as the distribution of (W_1, V_1) with $Z_1 = a$. Assume also that the sequence $\big((W_n, V_n'),\, n \geq 1\big)$ is independent of the i.i.d. sequence $(I_n,\, n \geq 1)$. Then $(W_1, V_1') \overset{\text{st}}{\leq} (W_1, V_1)$ if $Z_1 \geq a$, and so can put $\big((W_n, V_n, V_n'),\, n \geq 1\big)$ on the same probability space such that $V_n' \geq V_n$ for all $n \geq 1$. We now modify the Markov chain (3.23) as well, by defining $Z_1' = a$, and

$$Z_{n+1}' = \max\Big(Z_n' + W_n - r_1 V_n' - r_1 I_n, a\Big), \quad n \geq 1. \tag{3.24}$$

We then have $Z_n' \geq Z_n^* \geq Z_n$ for all $n \geq 1$. Defining

$$m_\gamma' = \inf\{n \geq 1 :\ Z_n' \geq \gamma\}$$

we see that $m_\gamma' \leq m_\gamma$. Therefore, (3.19) will follow once we prove that for some $\epsilon > 0$

$$\varliminf_{\lambda \to \infty} P\Big(m_\gamma' > \epsilon\gamma\Big(\gamma\overline{F}(\gamma)\Big)^{-k}\Big) > 0. \tag{3.25}$$

We now use Lemma 2. To this end we consider three systems. One is the present system we are considering (that is, we are admitting only sessions whose length does not exceed γ/K). The second one is the original system (we admit all sessions) and the last one is the system in which we admit only sessions whose length does not exceed $\gamma_0(K)/K$, where $\gamma_0(K)$ is the level defined in (3.22). We only consider $\gamma \geq \gamma_0(K)$. Observe that we can put all 3 systems on the same probability space in the following way. Start by generating an activity period of the third system. For each arrival in this activity period we always have an arrival in the first two systems, such that the corresponding session lengths for all three systems are ordered, with the longest for the second system, and the shortest for the third system. Furthermore, within the activity period of the third system we generate additional arrivals for the first system, and at the same time points we will also have an arrival for the original (second) system, whose length is at least the length of the corresponding session for the first system. Finally, still within the activity period of the third system we generate yet additional arrivals for the original (second) system. It is clear that the activity periods in the the first two systems will not end until the end of the activity period of the third system. Let V_1^* be the amount time within the activity period of the third system that the latter system is not empty, and let W_1^0 be the amount of work brought in the original (second) system within its activity period. Let $X = W_1^0 - r_1 V_1^* - r_1 I_1$. Let also $X^{(\gamma)} = W_1 - r_1 V_1' - r_1 I_1$, where W_1, I_1 and V_1' correspond to the system of interest (i.e. the first system). We assume that the state of all three systems at the time their activity periods start is a. Observe that with this construction we have $V_1' \geq V_1^*$, and $W_1 \leq W_1^0$. Therefore, $X^{(\gamma)} \leq X$, from which we conclude that $X^{(\gamma)} \overset{\text{st}}{\leq} X$. It follows from (3.22) that $EX < 0$. Furthermore, it follows from Theorem 1 of Resnick and Samorodnitsky (1997a) and the dominated variation of F that (2.2) holds. Finally, it follows from Proposition 1 of Resnick and Samorodnitsky (1997a) that for every $h > 0$ there is a K so large that

$$P\left(X^{(\gamma)} > \frac{1}{K^{1/2}}\gamma\right) \le P\left(W_1 > \frac{1}{K^{1/2}}\gamma\right) = o(\gamma^{-h})$$

as $\gamma \to \infty$. Thus we have verified all the conditions of Lemma 2, with $d = K^{1/2}$.

The claim (3.25) follows from Lemma 2 once we make sure that

$$\frac{h}{2} \ge 1 + k(\alpha - 1)$$

and

$$(\beta - 1)[K^{1/2}/3] \ge 1 + k(\alpha - 1),$$

where α and β are the Matuszewska indices in (1.6). However, both of these conditions will be satisfied once we choose K large enough. Therefore, we have proved that for K large enough (3.10) holds.

We now prove (3.9). We concentrate now on a system in which sessions arrive according to a Poisson process with rate $\lambda\overline{F}(\gamma/K)$. The service rate in the system is $k-1$. Let us denote by S_i the time of the arrival of the ith session, and let the events E_i be defined by (2.22) with $m = k-1$. Further, let ϵ_1 be a positive number satisfying

$$\epsilon_1 > \epsilon\lambda K^{\alpha/2}, \tag{3.26}$$

where α is the Matuszewska index from (1.6). We have by Lemma 4

$$\begin{aligned}
P\Big(\tau_\gamma^{(1)} \ge \epsilon\gamma\big(\gamma\overline{F}(\gamma)\big)^{-k}\Big) &\ge P\Bigg(\Big\{S_{\lceil\epsilon_1(\gamma\overline{F}(\gamma))^{-(k-1)}\rceil} \ge \epsilon\gamma\big(\gamma\overline{F}(\gamma)\big)^{-k}\Big\} \bigcap \bigcap_{j\le\epsilon_1(\gamma\overline{F}(\gamma))^{-(k-1)}} E_j\Bigg) \\
&\ge P\Big(S_{\lceil\epsilon_1(\gamma\overline{F}(\gamma))^{-(k-1)}\rceil} \ge \epsilon\gamma\big(\gamma\overline{F}(\gamma)\big)^{-k}\Big) - \Big(1 - P\Big(\bigcap_{j\le\epsilon_1(\gamma\overline{F}(\gamma))^{-(k-1)}} E_j\Big)\Big) \qquad (3.27) \\
&\ge P\Big(S_{\lceil\epsilon_1(\gamma\overline{F}(\gamma))^{-(k-1)}\rceil} \ge \epsilon\gamma\big(\gamma\overline{F}(\gamma)\big)^{-k}\Big) - \Big(1 - \big(P(E_1)\big)^{\epsilon_1(\gamma\overline{F}(\gamma))^{-(k-1)}}\Big).
\end{aligned}$$

Here $\lceil a\rceil$ is the smallest integer greater or equal to a. Observe that by (3.26) and the law of large numbers

$$P\Big(S_{\lceil\epsilon_1(\gamma\overline{F}(\gamma))^{-(k-1)}\rceil} \ge \epsilon\gamma\big(\gamma\overline{F}(\gamma)\big)^{-k}\Big) \to 1 \tag{3.28}$$

as $\gamma \to \infty$. Furthermore, by Lemma 6 there is a finite positive constant $c = c(K)$ such that for all γ big enough,

$$P(E_1) \ge 1 - c(K)\big(\gamma\overline{F}(\gamma)\big)^{k-1}.$$

Therefore, for all γ big enough,

$$1 - \big(P(E_1)\big)^{\epsilon_1(\gamma\overline{F}(\gamma))^{-(k-1)}} \le 1 - \frac{1}{2}e^{-\epsilon_1 c(K)}. \tag{3.29}$$

Now (3.9) follows from (3.27), (3.28) and (3.29), and the proof of the theorem is now complete. □

References

N. Bingham, C. Goldie and J. Teugels (1987): *Regular Variation.* Cambridge University Press.

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.*

R. Burton and E. Waymire (1985): Scaling limits for associated random measures. *Annals of Probability* 13:1267–1278.

N. Duffield and N. O'Connell (1995): Large deviations and overflow probabilities for the general single–server queue, with applications. *Mathematical Proceedings of Cambridge Philosophical Society* 118:363–374.

P. Embrechts and C. Goldie (1982): On convolution tails. *Stochastic Processes and Their Applications* 13:263–278.

P. Embrechts, C. Klüppelberg and T. Mikosch (1997): *Modelling Extreme Events for Insurance and Finance.* Springer-Verlag, Berlin.

A. Erramilli, O. Narayan and W. Willinger (1996): Experimental queueing analysis with long–range dependent packet traffic. *IEEE/ACM Transactions on Network Computing* 4:209–223.

J. Esary, F. Proschan and D. Walkup (1967): Association of random variables, with applications. *Ann. Math. Statist.* 38:1466–1474.

D. Heath, S. Resnick and G. Samorodnitsky (1996): Heavy tails and long range dependence in on/off processes and associated fluid models. To appear in *Mathematics of Operations Research.* Available as TR1144.ps.Z at http://www.orie.cornell.edu/trlist/trlist.html.

D. Heath, S. Resnick and G. Samorodnitsky (1997): Patterns of buffer overflow in a class of queues with long memory in the input stream. To appear in *the Annals of Applied Probability.* Available as patterns.ps at http://www.orie.cornell.edu/ gennady/techreports/imsjournals.

P. Jelenković and A. Lazar (1995): Subexponential asymptotics of a Markov–modulated $G/G/1$ queue. Technical report, Columbia University.

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.

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

S. Resnick (1992): *Adventures in Stochastic Processes.* Birkhäuser, Boston.

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

S. Resnick and G. Samorodnitsky (1997b): The effect of long range dependence in a simple queuing model. *Operations Research* 45.

B. Ryu and S. Lowen (1995): Modeling, analysis and simulation of self–similar traffic using the fractal–shot–noise–driven Poisson process. *Proceedings of IASTED International Conference on Modeling and Simulation*, Pittsburgh, PA.

S. Vamvakos and V. Anantharam (1996): On the departure process of a leaky bucket system with long–range dependent input traffic. Preprint.

W. Willinger, M. Taqqu, M. Leland and D. Wilson (1996): Self–similarity through high variability: statistical analysis of ethernet LAN traffic at the source level (extended version).

IEEE/ACM Transactions on Networking To appear.

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

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