The Annals of Applied Probability
1997, Vol. 7, No. 4, 1021–1057

PATTERNS OF BUFFER OVERFLOW IN A CLASS OF QUEUES WITH LONG MEMORY IN THE INPUT STREAM

BY DAVID HEATH,[1] SIDNEY RESNICK[1,2] AND GENNADY SAMORODNITSKY[1,2,3]

Cornell University

We study the time it takes until a fluid queue with a finite, but large, holding capacity reaches the overflow point. The queue is fed by an *on/off* process with a heavy tailed *on* distribution which is known to have long memory. It turns out that the expected time until overflow, as a function of capacity L, increases only polynomially fast; so overflows happen much more often than in the "classical" light tailed case, where the expected overflow time increases as an exponential function of L. Moreover, we show that in the heavy tailed case overflows are basically caused by single huge jobs. An implication is that the usual $GI/G/1$ queue with finite but large holding capacity and heavy tailed service times will overflow about equally often *no matter how much we increase the service rate*. We also study the time until overflow for queues fed by a superposition of k iid *on/off* processes with a heavy tailed *on* distribution, and we show the benefit of pooling the system resources as far as time until overflow is concerned.

1. Introduction. Traffic on data networks (e.g., Ethernet LAN's), has characteristics substantially different from those of traditional voice traffic. An important feature of data traffic lies in its dependence structure; traditional models are based on assumptions of short-range dependence (like Poisson arrivals and exponential call lengths), while recent measurement and analysis of data traffic has produced strong indications of long-range dependence and self-similarity. Several empirical studies present statistical evidence for existence of these nonstandard dependence structures. See, for example, Leland, Taqqu, Willinger and Wilson (1993, 1994), Willinger, Taqqu, Leland and Wilson (1995), Crovella and Bestavros (1995), and Cunha, Bestavros and Crovella (1995).

Seeking an explanation for the observed long-range dependence and self-similarity, Willinger, Taqqu, Sherman and Wilson (1995) have modeled traffic between a single source and destination as an *on/off* or *packet train* process. In their model, an idealized source alternates between an *on* state, in which it produces data at a constant rate, and an *off* state, in which it produces

Received August 1996; revised May 1997.

[1]Partially supported by NSA Grant MDA904-95-H-1036 at Cornell University.

[2]Received support from NSF Grant DMS-94-00535 at Cornell University.

[3]Also supported by United States–Israel Binational Science Foundation (BSF) Grant 92-00074.

AMS 1991 *subject classifications*. 60K25, 90B15.

Key words and phrases. Long range dependence, heavy tails, on/off models, $G/G/1$ queue, fluid models, long memory, heavy tailed distribution, regular variation, time to hit a level, buffer overflow, maximum work load, weak convergence.

no data. The durations of the on and off periods are independent; *on* times are identically distributed, and so are off times. The data they present indicate that both on and off times are reasonably well modeled by heavy tailed distributions with shape parameter governing heaviness represented by the parameter α. In one example, $\alpha = 1.7$ and 1.2, respectively, for the on and off periods. A similar conclusion was drawn by Crovella and Bestavros (1995), who in their study of World Wide Web use found evidence of heavy tails in such things as file lengths, transfer times and operator idle periods. Other papers dealing with on/off and related models for communication systems are Brichet, Roberts, Simonian and Veitch (1996), Kella and Whitt (1992) and Choudhury and Whitt (1995).

Various paradigms for on/off models can be kept in mind. One is the storage or fluid queue model where the store is filling at rate 1 during an on period and the contents are subject to constant release at rate r when the content level is positive. Another paradigm allows one to imagine work entering the system at rate 1 during on periods and a server working at rate r. We use either paradigm as is convenient.

In a previous paper we studied the stationary distributions of the simple on/off models. In the present paper we study the behavior of the first time the contents process exceeds level L for large levels L. Since this represents the time until "buffer overflow" in an on/off system with limited capacity, it is important in understanding the behavior of traffic networks.

The simplest model, consisting of a single on/off source feeding a single server queue, is defined as follows. Let $\{X_i, i = 1, 2, \ldots\}$ be a sequence of iid nonnegative random variables representing on periods, and similarly let $\{Y_i, i = 1, 2, \ldots\}$ be iid nonnegative random variables representing off periods. The on and off sequences are independent. Let F_{on} be the common distribution of X_i's, and let F_{off} be the common distribution of Y_i's. The work load arrives in the system at rate 1 during on periods (no work load arrives in the system during off periods). The service rate is r; that is, whenever the system is nonempty, work is leaving the system at rate r. The state of the system at time t (its content at time t, the work load in the system at time t) is denoted by $X(t)$ and can be formally defined as follows. For a $t \geq 0$ let

$$Z(t) = \begin{cases} 1, & \text{if } \sum_{i=1}^{n-1}(X_i + Y_i) \leq t < \sum_{i=1}^{n-1}(X_i + Y_i) + X_n, \\ & \quad \text{for some } n \geq 1, \\ 0, & \text{otherwise.} \end{cases} \tag{1.1}$$

So $Z(t)$ is the indicator of the the source "being on" at time t. Defining the service rate at state x by

$$r(x) = \begin{cases} r, & \text{if } x > 0, \\ 0, & \text{if } x = 0, \end{cases}$$

the state process $\{X(t), t \geq 0\}$ is defined by

$$dX(t) = Z(t)dt - r(X(t))\,dt. \tag{1.2}$$

The analogous $GI/G/1$ queue can be thought of as model (1.2) with on periods shrunk to zero and work load arriving in the system in lumps of size $\{B_i, i \geq 1\}$. In this context, $\{Y_i, i \geq 1\}$ can be thought of as interarrival times. One can take, for example, $B_i = (1-r)X_i$, as this is the net increase in the state of the system (1.2) after the ith on period. However, the discussion below does not depend on this particular form of the offered work. The service rate is still r, so that the actual service time of the ith customer is B_i/r, $i \geq 1$.

It turns out that the system overflow patterns in the on/off model, when the on times are heavy tailed, are very similar to those of the $GI/G/1$ queue when the amount of work B_i is heavy tailed. The computations describing the structure of system overflows in both cases are similar, and they are easier for the $GI/G/1$ queue. We will present the detailed arguments only for the more involved case of the fluid model (1.2).

To emphasize the dramatic effect of heavy tailed distributions on system behavior, we contrast the heavy tailed case with the simplest, classical case in Section 3. Consider the fluid queue with F_{on} being exponential with mean μ_{on}, F_{off} being exponential with mean μ_{off}, such that

$$\frac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}} < r. \tag{1.3}$$

Clearly, (1.3) is the necessary and sufficient condition for system (1.2) to be stable. If

$$\tau(L) = \inf\{t \geq 0 \colon X(t) \geq L\} \tag{1.4}$$

is the time until the system overflows, then based on martingale and Markov methods we find in Section 3 that the expected overflow time in this exponential case is

$$E\tau(L) = a\left(\exp\left(L\left(\frac{1}{(1-r)\mu_{\text{on}}} - \frac{1}{r\mu_{\text{off}}}\right)\right) - 1\right) - bL, \qquad L \geq 0, \tag{1.5}$$

where

$$a = \frac{(\mu_{\text{on}} + \mu_{\text{off}})(r\mu_{\text{off}})^2}{(r\mu_{\text{off}} - (1-r)\mu_{\text{on}})^2}$$

and

$$b = \frac{\mu_{\text{on}} + \mu_{\text{off}}}{r\mu_{\text{off}} - (1-r)\mu_{\text{on}}}.$$

Several conclusions are immediate from (1.5). First of all, in the exponentially distributed on and off periods case, the expected time till the system overflows increases exponentially fast with the system holding capacity. Second, the average off time μ_{off} critically affects the average time until the system overflows since it affects the multiplicative constant as well as the growth rate. Quite different conclusions are reached in the heavy tailed case.

In Section 2 we show that, in the case of on times having a heavy tailed distribution, the expected time to exceed L is asymptotically the same as the expected time until a single on period would cause the contents to exceed L, assuming the contents were empty at the start of the on time. This is very different from the exponential result (1.5). In this heavy tailed case, the expected time until a single on period causes the system to exceed the capacity L is asymptotic to ($\mu = \mu_{\text{on}} + \mu_{\text{off}}$)

$$\frac{\mu}{(1-r)^{\alpha}}\left(\frac{1}{1-F_{\text{on}}(L)}\right),$$

which is of smaller order of magnitude than (1.5). In the case of heavy tailed on periods the expected time until the system exceeds a level grows much slower than the exponential rate of increase seen in (1.5). Furthermore, the fact that in the heavy tailed case the system overflow is caused by a single long on period implies that the mean off time μ_{off} affects the expected time until overflow only by its effect on a multiplicative factor but does not otherwise influence the growth rate.

A similar conclusion is valid for the $GI/G/1$ queue with heavy tailed amounts of work $\{B_i, i \geq 1\}$. In this case the offered work load exceeds the system capacity L when a single customer brings amount of work reaching L. In particular, the mean interarrival time affects the time until the overflow only as a multiplicative factor, and it *does not depend on the service rate* r (!). This provides intuition about the "failure modes" of such a system.

Precise arguments showing unusual behavior in the heavy tailed case are presented in the next section, where we study the maximum of the fluid queue (1.2) over a single "wet period" and use the findings to obtain functional limit theorems for the maximum process of the queue (1.2) and for the hitting time process of the same queue. Section 3 contrasts in detail the behavior in cases where on and off distributions have exponentially bounded tails with that in the heavy tailed case. Tangentially relevent papers on extremes of queues (which typically emphasize Markovian methods and exponential tails) are Iglehart (1972), Asmussen and Perry (1992), Berger and Whitt (1995) and Abate, Choudhury and Whitt (1994).

In Section 4 we study the behavior of models with several on/off sources and a single server. We show again that in the case $r < 1$ the asymptotic behavior of the time at which the contents process exceeds L is the same as that of the first time that any of the input processes has an on period long enough to achieve level L from an empty initial content level. We then compare the behavior of a system of completely separate on/off processes with one in which the inputs are pooled and in which the capacity of the system is the sum of the capacities of the separate systems. Our conclusions quantify the benefits of pooling the system resources.

Other papers on multisource models, usually emphasizing Markovian environments, are Anick, Mitra and Sondhi (1982), Prabhu and Pacheco (1995) and Pacheco and Prabhu (1996).

2. Level crossing times in single input models. In this section we consider the extreme values of the contents process specified in (1.2) and the time for the content to cross a level. The fluid or storage model is generated by an alternating renewal process which feeds a reservoir. We represent the renewal sequence as $\{S_n, n \geq 0\}$ with $S_n = \sum_{i=1}^{n}(X_i + Y_i)$, $n \geq 1$, and for convenience we suppose $S_0 = 0$. Both F_{on} and F_{off} have finite means μ_{on} and μ_{off} and we set $\mu = \mu_{\text{on}} + \mu_{\text{off}}$. During an on period, liquid enters at net rate $1 - r$, and during an off period liquid is released at uniform rate r. We assure that neither the input rate nor the output rate overwhelms the other by assuming

$$1 > r > \frac{\mu_{\text{on}}}{\mu}. \tag{2.1}$$

Define $S_n^{(X)} = \sum_{i=1}^{n} X_i$ and $S_n^{(Y)} = \sum_{i=1}^{n} Y_i$ and the stopping time

$$\bar{N} = \inf\{n > 0\colon (1-r)S_n^{(X)} - rS_n^{(Y)} \leq 0\} \tag{2.2}$$

so that

$$\begin{aligned}[\bar{N} = n] &= \{(1-r)S_j^{(X)} - rS_j^{(Y)} > 0,\ j = 1, \ldots, n-1,\\ &\qquad (1-r)S_n^{(X)} - rS_n^{(Y)} \leq 0\}\\ &\in \mathscr{B}(X_i, Y_i, i = 1, \ldots, n).\end{aligned}$$

Consider $\{X(S_n), n \geq 0\}$. Comparing $X(S_n)$ with $X(S_{n+1})$ we get

$$\begin{aligned}X(S_{n+1}) &= (X(S_n) + (1-r)X_{n+1} - rY_{n+1})^+\\ &= (X(S_n) + \xi_{n+1})^+,\end{aligned} \tag{2.3}$$

where $\{\xi_{n+1} = (1-r)X_{n+1} - rY_{n+1}\}$ is iid. This equation expresses that the change of contents over a renewal interval is the input during the on period and the loss during the off period. Of course (2.3) is Lindley's equation [Resnick (1992), page 270; Asmussen (1988); Feller (1971)] and, since (2.1) implies

$$E\xi_1 = (1-r)\mu_{\text{on}} - r\mu_{\text{off}} = \mu_{\text{on}} - r\mu < 0,$$

we know from standard theory that the process

$$\{W_n\} := \{X(S_n)\}$$

will be stable and $E\bar{N} < \infty$. As is customary, we call $\{W_n\}$ the *queuing process*.

We suppose that

$$1 - F_{\text{on}}(x) = x^{-\alpha}L(x), \qquad \alpha > 1,\ x \to \infty, \tag{2.4}$$

where L is a slowly varying function. Note that the process $\{X(t), t \geq 0\}$ is regenerative [cf. Resnick (1992); Feller (1971); Asmussen (1988)]. One set of regeneration times is

$$\{C_n\} := \{S_n\colon X(S_n-) = 0\},$$

which are the times when a dry period ends and input commences to fill the store. In order to understand the behavior of the extremes of $\{X(t)\}$, it is natural to study the extremes over a cycle. For this purpose, it is necessary to understand the tail behavior of the distribution of maximum of the queuing process over one cycle. A result about this is stated next.

PROPOSITION 2.1. *For the stable queuing process $\{W_n\}$ satisfying (2.1) and (2.4), the maximum over a cycle has a distribution tail asymptotic to the tail of the on distribution; that is, as $x \to \infty$,*

$$\begin{aligned} P\Big[\bigvee_{n=0}^{\bar{N}} W_n > x\Big] &\sim P[\xi_1 > x]E(\bar{N}) \\ &\sim P[(1-r)X_1 > x]E(\bar{N}) \\ &\sim (1-r)^{\alpha}\bar{F}_{\text{on}}(x)E(\bar{N}). \end{aligned} \tag{2.5}$$

Note the result depends on F_{on} and r but that F_{off} only affects the answer through the multiplicative factor $E(\bar{N})$.

Independent and different proofs of this critical result have been given by the authors and by Asmussen (1998), who proves the result in the somewhat more general context of a random walk whose step distribution is subexponential. See Asmussen (1998) for details. Our original proof can be found at `http://www.orie.cornell.edu/~gennady/techreports/patterns.ps`. A nice review of the asymptotics of the tail of the all time maximum distribution of the random walk (as opposed to the cycle maximum) is given in Embrechts and Veraverbeke (1982).

We now look at the extremes of $\{X(t)\}$ over a cycle and examine the distribution tail of $\bigvee_{0\le s\le C_1} X(s)$, where

$$C_1 = S_{\bar{N}}.$$

Note that $\bar{N}$ is the first downgoing ladder epoch of the random walk

$$\left\{\sum_{i=1}^{n}\xi_i, n \ge 0\right\} = \{(1-r)S_n^{(X)} - rS_n^{(Y)}, n \ge 0\}$$

associated with the queuing process $\{W_n\}$ and that it is not the downgoing ladder epoch of $\{S_n\}$ which determines the time scale.

COROLLARY 2.2. *Assume the contents process $\{X(t)\}$ satisfies (2.1) and (2.4). The distribution tail of the maximum of the contents process over one cycle is asymptotic to the tail of the on distribution; that is, as $x \to \infty$,*

$$P\Big[\bigvee_{s=0}^{C_1} X(s) > x\Big] \sim (1-r)^{\alpha}\bar{F}_{\text{on}}(x)E(\bar{N}). \tag{2.6}$$

Note again that F_{off} only affects the answer through the multiplicative factor $E(\bar{N})$.

PROOF OF COROLLARY 2.2. Set $M_1 = \bigvee_{s=0}^{C_1} X(s)$. Because of the sawtooth character of the paths of $X(\cdot)$ we have that

$$M_1 = \bigvee_{j=1}^{\bar{N}} \left((1-r)S_j^{(X)} - rS_{j-1}^{(Y)}\right)$$

and therefore

$$M_1 \geq \bigvee_{j=0}^{\bar{N}} \left((1-r)S_j^{(X)} - rS_j^{(Y)}\right) =_d \bigvee_{n=0}^{\bar{N}} W_n.$$

Thus

$$\begin{aligned}\liminf_{x\to\infty} \frac{P[M_1 > x]}{\bar{F}_{\text{on}}(x)} &\geq \liminf_{x\to\infty} \frac{P[\bigvee_{j=0}^{\bar{N}} W_n > x]}{\bar{F}_{\text{on}}(x)} \\ &= (1-r)^{\alpha} E(\bar{N}),\end{aligned}$$

where the last step uses Proposition 2.1.

To get a reverse inequality, choose K such that

$$E((1-r)X_1 - r(Y_1 \wedge K)) < 0,$$

which can always be done since

$$E(Y_1 \wedge K) \uparrow EY_1$$

as $K \uparrow \infty$. Then

$$S_j^{(Y,K)} := \sum_{i=1}^{j} (Y_i \wedge K) \leq S_j^{(Y)},$$

which obviously gives

$$\bigvee_{j=0}^{\bar{N}} \left((1-r)S_j^{(X)} - rS_{j-1}^{(Y)}\right) \leq \bigvee_{j=1}^{\bar{N}} \left((1-r)S_j^{(X)} - r\,S_{j-1}^{(Y,K)}\right).$$

Also

$$\bar{N} \leq \bar{N}^{(K)} := \inf\{n > 0\colon (1-r)S_n^{(X)} - rS_n^{(Y,K)} \leq 0\}$$

and thus we have

$$\begin{aligned}M_1 &\leq \bigvee_{j=0}^{\bar{N}^{(K)}} \left((1-r)S_j^{(X)} - rS_{j-1}^{(Y,K)}\right) \\ &\leq \bigvee_{j=0}^{\bar{N}^{(K)}} \left((1-r)S_j^{(X)} - rS_j^{(Y,K)} + r(Y_j \wedge K)\right) \\ &\leq \bigvee_{j=0}^{\bar{N}^{(K)}} \left((1-r)S_j^{(X)} - rS_j^{(Y,K)} + rK\right).\end{aligned}$$

We therefore have

$$\limsup_{x\to\infty} \frac{P[M_1 > x]}{\bar{F}_{\text{on}}(x)} \le \limsup_{x\to\infty} \frac{P\left[\bigvee_{j=0}^{\bar{N}^{(K)}} \left((1-r)S_j^{(X)} - rS_j^{(Y,K)}\right) > x - rK\right]}{\bar{F}_{\text{on}}(x)}$$

and applying Proposition 2.1 to the random walk $\{(1-r)S_j^{(X)} - rS_j^{(Y,K)},\ j \ge 0\}$ we get this equal to

$$(1-r)^\alpha E(\bar{N}^{(K)}) \to (1-r)^\alpha E(\bar{N})$$

as $K \to \infty$. This provides the reverse inequality and completes the proof. □

We are now in a position to discuss the behavior of the extremes of the contents process and also the behavior of the first passage time over a level. For a nondecreasing function $U\colon (0,\infty) \mapsto (0,\infty)$ define the (left continuous) inverse

$$U^{\leftarrow}(x) = \inf\{s > 0\colon U(s) \ge x\}, \qquad x > 0.$$

Define the nondecreasing process

$$M(t) = \bigvee_{s=0}^{t} X(s)$$

and the first passage time ($L > 0$)

$$\begin{aligned}\tau(L) &= \inf\{s > 0\colon X(s) \ge L\}\\ &= \inf\{s > 0\colon M(s) \ge L\}\\ &= M^{\leftarrow}(L).\end{aligned}$$

Standard inversion techniques from extreme value theory [Resnick (1986, 1987), Section 4.4] allow for the simultaneous treatment of the weak convergence properties of $(M(\cdot), M^{\leftarrow}(\cdot))$ as random elements of $D_r(0,\infty)\times D_l(0,\infty)$, where $D_r(0,\infty)$ is the space of right continuous functions with finite left limits and $D_l(0,\infty)$ is the space of left continuous functions on $(0,\infty)$ with finite right limits. Each space is equipped with the M_1-topology [Avram and Taqqu (1986, 1989, 1992)].

THEOREM 2.3. *Assume the contents process $\{X(t)\}$ satisfies (2.1) and (2.4). Define the quantile function*

$$b(s) = \left(\frac{1}{1-F_{\text{on}}}\right)^{\leftarrow}(s).$$

Let $\{Y_\alpha(t), t > 0\}$ be the extremal process [Resnick (1987), Section 4.3) generated by the extreme value distribution

$$\Phi_\alpha(x) = \exp\{-x^{-\alpha}\}, \qquad x > 0$$

so that

$$P[Y_\alpha(t) \le x] = \Phi_\alpha^t(x).$$

Define

$$S_\alpha(t) = \frac{1-r}{\mu^{1/\alpha}} Y_\alpha(t).$$

Then, in $D_r(0,\infty) \times D_l(0,\infty)$ *as* $u \to \infty$,

$$\left(\frac{M(u\cdot)}{b(u)}, \left(\frac{M(u\cdot)}{b(u)} \right)^{\leftarrow} \right) \Rightarrow (S_\alpha, S_\alpha^{\leftarrow}).$$

In particular we get for the first passage process, as $u \to \infty$,

$$(1 - F_{\text{on}}(u))\tau(u\cdot) \Rightarrow Y_\alpha^{\leftarrow}\left(\frac{\mu^{1/\alpha}}{1-r} \cdot \right)$$

and

$$\lim_{L\to\infty} P\left[\frac{(1-r)^\alpha}{\mu}(1 - F_{\text{on}}(L))\tau(L) \le x \right] = P[E(1) \le x] = 1 - e^{-x}, \qquad x > 0,$$

where $E(1)$ *is a unit exponential random variable. Furthermore, as* $L \to \infty$,

$$(1 - F_{\text{on}}(L))E(\tau(L)) \to \frac{\mu}{(1-r)^\alpha}.$$

PROOF. We let $\{\bar{N}_k, k \ge 1\}$ be the iterates of $\bar{N}$ so that $\bar{N}_k$ is the kth downgoing ladder epoch of the random walk $\{(1-r)S_n^{(X)} - rS_n^{(Y)}, n \ge 0\}$. Then by the strong law of large numbers $\bar{N}_k/k \to E(\bar{N})$ as $k \to \infty$. We write

$$M(S_{\bar{N}_k}) = \bigvee_{s=0}^{S_{\bar{N}_k}} X(s) = \bigvee_{i=1}^{k} \left(\bigvee_{s=S_{\bar{N}_{i-1}}}^{S_{\bar{N}_i}} X(s) \right) := \bigvee_{i=1}^{k} M_i$$

so that $\{M_i, i \ge 1\}$ is iid. From Corollary 2.2, as $x \to \infty$,

$$P[M_1 > x] \sim (1-r)^\alpha \bar{F}_{\text{on}}(x) E(\bar{N})$$

so that, as $u \to \infty$,

$$\begin{aligned} uP[M_1 > b(u)x] &\sim (1-r)^\alpha E(\bar{N}) u \bar{F}_{\text{on}}(b(u)x) \\ &\sim (1-r)^\alpha E(\bar{N}) x^{-\alpha}. \end{aligned}$$

Therefore,

$$\bigvee_{i=1}^{[ut]} \frac{M_i}{b(u)} = \frac{M(S_{\bar{N}_{[ut]}})}{b(u)} \Rightarrow (1-r)(E\bar{N})^{1/\alpha} Y_\alpha(t). \tag{2.7}$$

Observe that as, $u \to \infty$,

$$\frac{S_{\bar{N}_{[ut]}}}{u} \to \mu E(\bar{N}) t \tag{2.8}$$

in $C(0,\infty)$. For the renewal sequence $\{S_{\bar{N}_k}, k \ge 0\}$, let

$$\Theta(t) = \inf\{k\colon S_{\bar{N}_k} \ge t\}$$

be the associated counting function so that, as $u \to \infty$,

$$\frac{\Theta(ut)}{u} \to \frac{t}{ES_{\tilde{N}}} = \frac{t}{\mu E(\tilde{N})} \tag{2.9}$$

in $C(0, \infty)$. Note the inequalities

$$\frac{M(S_{\tilde{N}_{\Theta(ut)-1}})}{b(u)} \le \frac{M(ut)}{b(u)} \le \frac{M(S_{\tilde{N}_{\Theta(ut)}})}{b(u)}.$$

Now from Billingsley [(1968), Theorem 4.4] and composition

$$\begin{aligned}\frac{M(S_{\tilde{N}_{\Theta(ut)}})}{b(u)} = \frac{M(S_{\tilde{N}_{[u\cdot\Theta(ut)/u]}})}{b(u)} &\Rightarrow (1-r)(E\tilde{N})^{1/\alpha} Y_\alpha\left(\frac{t}{\mu E\tilde{N}}\right) \\ &=_d (1-r)\mu^{-1/\alpha} Y_\alpha(t) =: S_\alpha(t)\end{aligned}$$

in the J_1-topology, and we hope the same result is true in the M_1-topology for the family of processes $M(u\cdot)/b(u)$ as $u \to \infty$. In order to verify this, we need to show

$$\frac{M(S_{\tilde{N}_{\Theta(ut)}}) - M(S_{\tilde{N}_{\Theta(ut)-1}})}{b(u)} \Rightarrow 0 \tag{2.10}$$

in the M_1-topology. For a fixed t we get for any ε and large u that

$$\begin{aligned}P\left[\left|\frac{M(S_{\tilde{N}_{\Theta(ut)}}) - M(S_{\tilde{N}_{\Theta(ut)-1}})}{b(u)}\right| > \eta\right] &\le P\left[\left|\frac{M(S_{\tilde{N}_{\Theta(ut)}}) - M(S_{\tilde{N}_{\Theta(u(t-\varepsilon))}})}{b(u)}\right| > \eta\right] \\ &\to P[|S_\alpha(t) - S_\alpha(t-\varepsilon)| > \eta],\end{aligned}$$

which goes to 0 as $\varepsilon \to 0$ by the stochastic continuity of S_α [Resnick (1987), Proposition 4.7). The multivariate analogue needed to prove M_1-convergence is similar.

The weak convergence result for $\tau(\cdot)$ is obtained by taking inverses in the process convergence. Inversion is a continuous operation in the M_1-topology. We note that inverses of extremal processes have exponential marginals [Resnick (1986, 1987)] so, as $u \to \infty$,

$$\frac{M^{\leftarrow}(b(u)x)}{u} \Rightarrow S_\alpha^{\leftarrow}(x),$$

and changing variables $s \mapsto b(u)$ yields

$$\frac{M^{\leftarrow}(sx)}{1/(1-F_{\mathsf{on}}(s))} \Rightarrow S_\alpha^{\leftarrow}(x).$$

Observe, for $y > 0$,

$$\begin{aligned}P[S_\alpha^{\leftarrow}(1) \le y] &= P[1 \le S_\alpha(y)] \\ &= P\left[\frac{\mu^{1/\alpha}}{1-r} \le Y_\alpha(y)\right] \\ &= 1 - \exp\{-y(1-r)^\alpha \mu^{-1}\}.\end{aligned}$$

Finally we consider the result for the expected values. On the one hand, by Fatou's lemma, we get

$$1 \le \liminf_{L\to\infty} E\left(\frac{(1-r)^{\alpha}}{\mu}(1-F_{\text{on}}(L))\tau(L)\right).$$

For a reverse inequality, note that

$$\tau(L) \le S_{\nu},$$

where

$$\nu := \inf\{n\colon X_n > L/(1-r)\}$$

so that

$$E\tau(L) \le E(X_1+Y_1)E\nu = \mu E\nu.$$

However,

$$\begin{aligned} E\nu &= \sum_{n=0}^{\infty} P[\nu > n] = \sum_{n=0}^{\infty} P\left[\bigvee_{i=1}^{n} X_i \le \frac{L}{1-r}\right] \\ &= \frac{1}{1-F_{\text{on}}(L/(1-r))} \sim \frac{(1-r)^{-\alpha}}{\bar{F}_{\text{on}}(L)} \end{aligned}$$

and this completes the proof. □

To illustrate these results we present two modest simulations. For each simulation we supposed F_{on} was Pareto with $\alpha = 1.5$ and $r = 0.53$. For the first simulation (Figure 1), F_{off} was the same Pareto; for the second simulation (Figure 2), F_{off} corresponded to constant off times with value 3. We used 500 replications to compute expected hitting times of various levels by simulation and compared these with the approximate mean hitting time given by Theorem 2.3. The levels used for both experiments were 2, 5, 10, 22, 46, 100, 215, 464. The plots use a log scale for both axes. Note that the dotted line appears closer to the solid one when the off time is deterministic, which may indicate a faster rate of convergence of the the approximation compared to the situation where the approximation has to cope with randomness in the off time. However, no systematic investigation has been completed of the rate of convergence.

We also sought experimental evidence to confirm the intuition that in the heavy tailed case the process exceeds a level L because of a very long on period. As an additional experiment, we simulated 1000 runs of the process with $\alpha = 1.5$, the off distribution concentrated at 3 and $r = 0.53$. We waited until the process crossed $L = 64$ and then measured the length of the last on period X_{on}, multiplying by $(1-r)$. We compiled 1000 realizations of

$$\left(\frac{(1-r)X_{\text{on}}}{L}\right) \bigwedge 3,$$

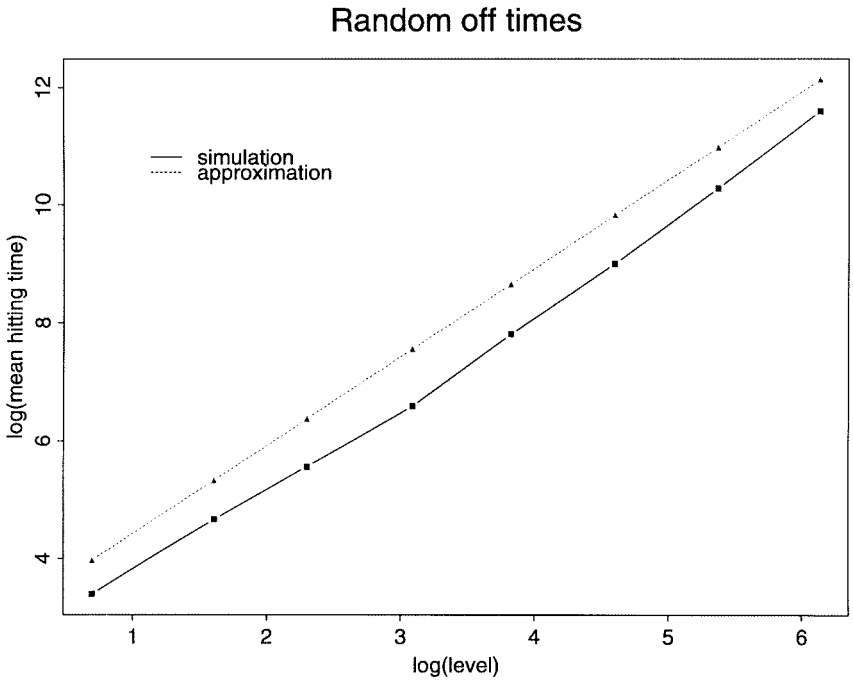

FIG. 1. *Pareto on/off periods,* $\alpha = 1.5$, $r = 0.53$.

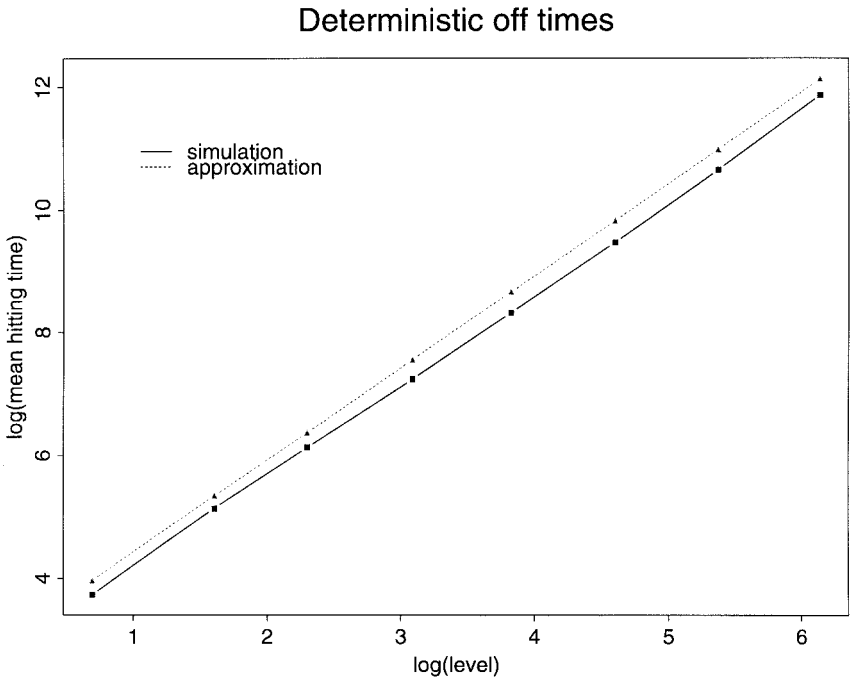

FIG. 2. *Pareto on period, deterministic off period,* $\alpha = 1.5$, $r = 0.53$.

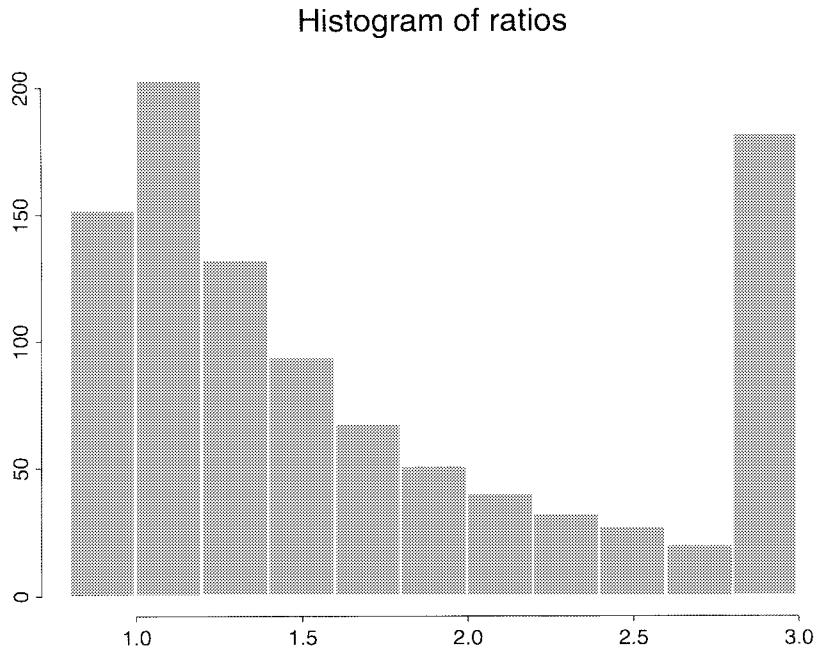

FIG. 3. $((1-r)X_{\text{on}}/L) \wedge 3$ *for Pareto on period, deterministic off period,* $\alpha = 1.5$, $r = 0.53$.

the truncation by 3 being for the purpose of keeping the data in a comfortable range. The range of the 1000 realizations was [0.896, 3] and 848 observations were at least as large as 1, meaning that in about 85% of the simulation runs the process crossed L due to a single large on period pushing the process across. Figure 3 is a histogram of the data, showing the preponderance of observations to the right of 1.

3. Contrast with exponential tails. If F_{on} has an exponentially bounded tail, known results of Iglehart (1972) can be applied to obtain the analogue of Theorem 2.3. We continue to assume that $\{X_i\}$ and $\{Y_i\}$ are iid, independent of each other, with distributions F_{on} and F_{off}, respectively. Set $\xi_i = (1-r)X_i - rY_i$ and for stability continue to suppose $E\xi_1 < 0$. Recall $\bar{N}$ is the first downgoing ladder epoch of the random walk with steps $\{\xi_i\}$. We need to suppose the following:

A1. for some $\gamma > 0$, $E^{\gamma\xi_1} = 1$;

for this value of γ we have the following:

A2. $E\xi_1 \exp(\gamma\xi_1) := \mu_\gamma \in (0, \infty)$;
A3. ξ_1 has a nonlattice distribution.

An analogue of Corollary 2.2 is provided by Iglehart (1972) which suffices to give the tail behavior of the maximum contents level in a cycle. We write $S_n^{(\xi)} = (1-r)S_n^{(X)} - rS_n^{(Y)}$, $n \geq 0$.

PROPOSITION 3.1 [Iglehart (1972), Lemma 4]. *Suppose assumptions* A1, A2 *and* A3 *hold. Then, for* $x > 0$,

$$P[M_1 > x] = P\Big[\bigvee_{s=0}^{C_1} X(s) > x\Big] \sim a(0)e^{-\gamma x}, \tag{3.1}$$

where

$$a(0) = \frac{\big(1 - E(\exp(\gamma S_{\bar{N}}^{(\xi)}))\big)^2}{\gamma\mu_\gamma E(\bar{N})} E\big(\exp(\gamma(1-r)X_1)\big).$$

We may now follow the line of reasoning of Section 2. The exponential tails given in (3.1) imply

$$\gamma \bigvee_{i=1}^{[ut]} M_i - \log a(0)u \Rightarrow Y_0(t) \tag{3.2}$$

in $D_r(0,\infty)$, where $Y_0(\cdot)$ is the extremal process generated by the Gumbel distribution

$$\Lambda(x) = \exp\{-e^{-x}\}, \qquad x \in \mathbb{R}.$$

We then get, as $u \to \infty$,

$$\begin{aligned}&\big(\gamma M(ut) - \log a(0)u,\, M^{\leftarrow}((x + \log a(0)u)/\gamma)/u\big)\\ &\quad\Rightarrow (Y_0(t/(\mu E(\bar{N})), \mu E(\bar{N}) Y_0^{\leftarrow}(x)),\end{aligned}$$

which leads to

$$a(0)\frac{\tau(x+u)}{e^{\gamma u}} \Rightarrow \mu E(\bar{N}) Y_0^{\leftarrow}(\gamma x).$$

If $x = 0$, we get, as $u \to \infty$,

$$a(0)\frac{\tau(u)}{e^{\gamma u}} \Rightarrow \mu E(\bar{N}) Y_0^{\leftarrow}(0).$$

Note that $Y_0^{\leftarrow}(0)$ is exponentially distributed with mean 1, and we get the final result

$$\begin{aligned}\left(\frac{a(0)}{\mu E(\bar{N})}\frac{\tau(u)}{\exp(\gamma u)}\right) &= \left(E\big(\exp(\gamma(1-r)X_1)\big)\frac{(1 - E\exp(-\gamma S_{\bar{N}}^{(\xi)}))^2}{\mu(E\bar{N})^2\gamma\mu_\gamma}\right)\\ &\quad\times \frac{\tau(u)}{\exp(\gamma u)} \Rightarrow E(1)\end{aligned} \tag{3.3}$$

as $u \to \infty$, where $E(1)$ is a unit exponential random variable.

We may also check that $E\tau(u)/e^{\gamma u}$ converges as follows. Observe that

$$0 \le \tau(s) \le S_{\tilde{N}_{V(s)}}, \tag{3.4}$$

where

$$V(s) = \inf\left\{n\colon \bigvee_{i=1}^{n} M_i \ge s\right\}$$

since the hitting time of $X(\cdot)$ must occur before the end of the cycle that has a cycle maximum bigger than the level. Since $V(s)$ is geometrically distributed

$$a(0)\frac{V(s)}{e^{\gamma s}} \Rightarrow E(1),$$

where, as usual, $E(1)$ is a unit exponential random variable. Furthermore, as $s \to \infty$,

$$\begin{aligned} ES_{\tilde{N}_{V(s)}} &= \mu E(\tilde{N}_{V(s)}) = \mu E(\tilde{N})E(V(s)) \\ &= \mu E(\tilde{N})\left(\frac{1}{P[M_1 > s]}\right) \\ &\sim \frac{\mu E(\tilde{N})}{a(0)}e^{\gamma s}. \end{aligned}$$

So, as $s \to \infty$,

$$\frac{E(S_{\tilde{N}_{V(s)}})}{e^{\gamma s}} \to \frac{\mu E(\tilde{N})}{a(0)}$$

and

$$\frac{S_{\tilde{N}_{V(s)}}}{e^{\gamma s}} \sim \mu E(\tilde{N})\frac{V(s)}{e^{\gamma s}} \Rightarrow \frac{\mu E(\tilde{N})}{a(0)}E(1).$$

These two statements coupled with (3.3) and (3.4) and a variant of Fatou's lemma sometimes called Pratt's lemma [Pratt (1960)] yield the desired result

$$E\tau(s) \sim \frac{\mu E(\tilde{N})}{a(0)}e^{\gamma s}, \qquad s \to \infty. \tag{3.5}$$

Note how critically F_{off} enters into formulas (3.3) and (3.5) since F_{off} is important in determining the growth rate γ of the hitting time. In the heavy tail case, F_{off} did not play a role in determining the growth rate of $\tau(s)$ since levels were hit basically due just to big upward jumps which were controlled by F_{on}. Recall that in the heavy tailed case, as $s \to \infty$,

$$E\tau(s) \sim \mu(1-r)^{-\alpha}/\bar{F}_{\text{on}}(s) \sim ES_\nu = \mu E\nu,$$

where

$$\nu = \inf\{n\colon (1-r)X_n > s\}.$$

EXAMPLE. Consider the standard example where both F_{on} and F_{off} are exponential distributions with means μ_{on} and μ_{off}, respectively. The negative drift condition is

$$(1-r)\mu_{\text{on}} - r\mu_{\text{off}} < 0,$$

and γ must satisfy

$$1 = E\exp(\gamma((1-r)X_1 - rY_1)) = \big((1-\gamma(1-r)\mu_{\text{on}})(1+\gamma r\mu_{\text{off}})\big)^{-1}.$$

Solving for γ we get the solutions $\gamma = 0$ and

$$\gamma = \frac{r\mu_{\text{off}} - (1-r)\mu_{\text{on}}}{(1-r)r\mu_{\text{on}}\mu_{\text{off}}} = \frac{-E\xi_1}{(1-r)r\mu_{\text{on}}\mu_{\text{off}}}.$$

The numerator is positive by the drift condition.

We now calculate the coefficient of $\tau(u)/e^{\gamma u}$ in (3.3) in order to compare it to the exact calculation given in (1.5). Since X_1 is assumed exponential with mean μ_{on} we have

$$E\exp(\gamma(1-r)X_1) = \frac{r\mu_{\text{off}}}{(1-r)\mu_{\text{on}}}.$$

To calculate μ_γ we compute $(d/ds)Ee^{s\xi_1}$ and substitute $s = \gamma$. This calculation is made easier by use of the formulas

$$1 + r\gamma\mu_{\text{off}} = \frac{r\mu_{\text{off}}}{(1-r)\mu_{\text{on}}},$$

$$1 - \gamma(1-r)\mu_{\text{on}} = \frac{(1-r)\mu_{\text{on}}}{r\mu_{\text{off}}} = \frac{1}{1+r\gamma\mu_{\text{off}}}.$$

With these formulas we find

$$\mu_\gamma = -E\xi_1.$$

Next observe that, because of exponential tails,

$$S^{(\xi)}_{\bar N} =_d -r\mu_{\text{off}}E(1),$$

where $E(1)$ is a unit exponential random variable and hence

$$\begin{aligned} 1 - E\exp(\gamma S^{(\xi)}_{\bar N}) &= 1 - \frac{1}{1+\gamma r\mu_{\text{off}}} \\ &= 1 - \frac{(1-r)\mu_{\text{on}}}{r\mu_{\text{off}}} \\ &= \frac{-E\xi_1}{r\mu_{\text{off}}}. \end{aligned}$$

Knowing the distribution of $S^{(\xi)}_{\bar N}$ also enables us to compute $E\bar N$ since

$$ES^{(\xi)}_{\bar N} = -r\mu_{\text{off}} = E(\bar N)E\xi_1$$

and hence

$$E\bar{N} = \frac{r\mu_{\text{off}}}{-E\xi_1}.$$

Putting the ingredients together, (3.2) becomes

$$\frac{(-E\xi_1)^2}{\mu(r\mu_{\text{off}})^2}\frac{\tau(u)}{e^{\gamma u}} \Rightarrow E(1),$$

which agrees with the exact calculation for the expected value in (1.5).

The contrast with the heavy tailed case is very evident. Instead of $E\tau(s)$ being of the same order as ES_ν, the expected time until an on period of length at least $s/(1-r)$ occurs, we have in our exponential example

$$\begin{aligned} ES_\nu &= \mu E\nu = \mu/\bar{F}_{\text{on}}(s/(1-r)) \\ &= \mu \exp\{s/(\mu_{\text{on}}(1-r))\}. \end{aligned}$$

However, from (3.5),

$$E\tau(s) \sim \frac{\mu(r\mu_{\text{off}})^2}{(-E\xi_1)^2}e^{\gamma s}.$$

Comparing the growth rates γ with $1/(\mu_{\text{on}}(1-r))$ we have

$$0 < \gamma = \frac{1}{(1-r)\mu_{\text{on}}} - \frac{1}{r\mu_{\text{off}}} < \frac{1}{(1-r)\mu_{\text{on}}}$$

so that ES_ν has a faster growth rate, which is to be expected since in the exponential case the process $X(\cdot)$ jumps over a high level as a result of an accumulation of small upward movements and not typically as a result of a single large jump.

To obtain the exact expression for $E\tau(s)$ in this example, proceed as follows. Defining

$$\tilde{X}(t) = (X(t), Z(t)), \qquad t \geq 0,$$

we describe the state of the system prior to reaching level s as a Markov process $\{\tilde{X}(t), t \geq 0\}$ with state space $\mathbb{E} = \{(x, i), 0 \leq x \leq s, i = 0, 1\}$. We can express $\tau(s)$ in terms of the hitting times of the Markov process $\{\tilde{X}(t), t \geq 0\}$ as

$$\tau(s) = T_{(s,1)} := \inf\{t \geq 0\colon \tilde{X}(t) = (s, 1)\}.$$

For x $\in \mathbb{E}$, let H(x) be the expected hitting time $T_{(s,1)}$ starting at x, and define, for an $0 \leq x \leq s$,

$$h_1(x) = H((x, 1)), \qquad h_2(x) = H((x, 0)).$$

Then $E\tau(s) = h_1(0)$. Using the natural filtration $\mathcal{F}_t = \sigma(Z(u), 0 \leq u \leq t)$, $t \geq 0$, we observe that, for any $t \geq 0$,

$$E(T_{(s,1)}|\mathcal{F}_t) = H\big(\tilde{X}(t \wedge T_{(s,1)})\big) + \tilde{X}(t \wedge T_{(s,1)}) := M(t).$$

Therefore, $\{M(t), t \geq 0\}$ is a martingale, and its martingale property leads to the following system of ordinary differential equations:

$$(1-r)h_1'(x) = -1 + \frac{1}{\mu_{\text{on}}}h_1(x) - \frac{1}{\mu_{\text{on}}}h_2(x), \tag{3.6}$$

$$rh_2'(x) = 1 + \frac{1}{\mu_{\text{off}}}h_1(x) - \frac{1}{\mu_{\text{off}}}h_2(x), \tag{3.7}$$

with the obvious boundary conditions

$$h_2(0) = \mu_{\text{off}} + h_1(0), \quad h_1(s) = 0. \tag{3.8}$$

The system (3.6)–(3.8) can be solved in the standard way, and we obtain (1.5).

4. A single server fluid queue fed by several on/off processes. Let $\{X_i^{(j)}, i = 1, 2, \ldots\}$, $j = 1, \ldots, k$, and $\{Y_i^{(j)}, i = 1, 2, \ldots\}$, $j = 1, \ldots, k$, be iid copies of the on sequence $\{X_i, i = 1, 2, \ldots\}$ and the off sequence $\{Y_i, i = 1, 2, \ldots\}$ correspondingly. We construct k iid on/off processes $\{Z_j(t), t \geq 0\}$, $j = 1, \ldots, k$, as in (1.1). Sometimes we will find it convenient to work with stationary versions of $\{Z_j(t), t \geq 0\}$, $j = 1, \ldots, k$. Those exist due to the finiteness of μ_{on} and μ_{off}, and can be constructed as follows. Fix a $j = 1, \ldots, k$, and let $C_j^{(0)}$, $X_j^{(0)}$, $Y_j^{(0)}$ and $Y_0^{(j)}$ be four independent random variables which are independent of $\{X_i^{(j)}, Y_i^{(j)}, i \geq 1\}$ defined as follows: $C_j^{(0)}$ is a Bernoulli random variable with values $\{0, 1\}$ and mass function

$$P[C_j^{(0)} = 1] = \frac{\mu_{\text{on}}}{\mu} = 1 - P[C_j^{(0)} = 0]$$

and ($x > 0$)

$$P[X_j^{(0)} > x] = \int_x^\infty \frac{\bar{F}_{\text{on}}(s)}{\mu_{\text{on}}}\, ds, \qquad P[Y_j^{(0)} > x] = \int_x^\infty \frac{\bar{F}_{\text{off}}(s)}{\mu_{\text{off}}}\, ds.$$

Finally, $Y_0^{(j)}$ has the F_{off} distribution. Define a delay random variable $D_j^{(0)}$ by

$$D_j^{(0)} = C_j^{(0)}(X_j^{(0)} + Y_0^{(j)}) + (1 - C_j^{(0)})Y_j^{(0)}$$

and a delayed renewal sequence by

$$\{S_n^{(j)}, n \geq 0\} := \left\{D_j^{(0)}, D_j^{(0)} + \sum_{i=1}^n (X_i^{(j)} + Y_i^{(j)}), n \geq 1\right\}. \tag{4.1}$$

Then a stationary version of $\{Z_j(t), t \geq 0\}$ is defined by

$$Z_j(t) = C_j^{(0)} 1_{[0,\, X_j^{(0)})}(t) + \sum_{n=0}^\infty 1_{[S_n^{(j)} \leq t < S_n^{(j)} + X_{n+1}^{(j)}]}. \tag{4.2}$$

See Heath, Resnick and Samorodnitsky (1996) for details. In a similar way we can construct a stationary version $\{Z_j(t), < -\infty < t < \infty\}$ defined for all real t. We take, further, the k stationary on/off processes to be independent.

In this section we consider a single server fluid queue, with service rate r, fed by k on/off processes. The combined inflow rate is given by

$$Z^{(k)}(t) = Z_1(t) + \cdots + Z_k(t), \qquad t \geq 0 \text{ or } -\infty < t < \infty, \tag{4.3}$$

and, similarly to (1.2), the state $\{X^{(k)}(t), t \geq 0\}$ of the system satisfies

$$dX^{(k)}(t) = Z^{(k)}(t)\,dt - r(X^{(k)}(t))\,dt. \tag{4.4}$$

It is of interest to consider the behavior of a system (4.4) with a general k, first of all as a step toward understanding queues with more general long memory input streams and, second, to understand the effect of pooling resources in the systems of the type we are considering. The natural rate condition for this system, parallel to (1.3), is

$$k\frac{\mu_{\text{on}}}{\mu} < r, \tag{4.5}$$

saying that the long-term inflow rate to the system (4.4) is less than the potential outflow rate. Of course, we also assume that $r < k$, to make sure that the system is nondegenerate. Although we do not enter into details here, we can verify that under condition (4.5) there is indeed a stationary stochastic process $\{X^{(k)}(t), t \geq 0\}$ satisfying (4.4). We assume, as always, that the distribution $F_{\text{on}} * F_{\text{off}}$ is not arithmetic.

When the on periods have a heavy tailed distribution, we know from the discussion in Section 2 that, for $k = 1$, the state of a system driven by (4.4) crosses a high level L by increasing to that level almost from 0 within a single on period. We expect high level crossing patterns of the system contents to be similar for a general k. Intuitively, the time to reach a high level L should critically depend on

$$k_0 = \text{the smallest integer} > r. \tag{4.6}$$

By the nontriviality assumption, $k_0 \leq k$. If $k_0 = 1$, by analogy to the case $k = 1$ one expects the state of the system to reach a high level L when one of the k on/off processes has an on period of length $L/(1 - r)$. If $k_0 > 1$, the same intuition says that the system will reach a high level L only when k_0 very long on periods occur at about the same time, and so it will take much longer until this high level is reached. In this section we prove the above statement for the case $k_0 = 1$, thus generalizing the conclusion reached in Section 2 for $k = 1$. The proof of Theorem 4.1 is significantly more involved than the argument required for $k = 1$ due, once again, to the lack of renewal structure in the process $\{Z^{(k)}(t), t \geq 0\}$.

A natural way of calculating the time until the system contents reach the level L is by starting from the moment the system is empty, and all k on/off processes begin an on period. One must realize, however, that for $k > 1$ such a moment in time is far from "typical" and, even if we initialize the system in such a way, chances are that such moments will not recur. Therefore, we state our theorem in a more general way, by allowing more general initial conditions. To this end, let H be an arbitrary probability law on $\mathbb{R}_+^k$ whose marginals have

finite first moments. Let $(D_1^{(0)}, \ldots, D_k^{(0)})$ be an H-distributed random vector, independent of $\{X_i^{(j)}, Y_i^{(j)}, i \geq 1, 1 \leq j \leq k\}$. We again define a delayed renewal sequence by (4.1), and, similarly to (4.2), we define $\{Z_j(t), t \geq 0\}$ by

$$Z_j(t) = \sum_{n=0}^{\infty} 1_{[S_n^{(j)} \leq t < S_n^{(j)} + X_{n+1}^{(j)}]}. \tag{4.7}$$

Clearly, this time $\{Z_j(t), t \geq 0\}$ does not have to be stationary. If $\{X^{(k)}(t), t \geq 0\}$ is given now by $X^{(k)}(0) = x_0 \in [0, \infty)$ and (4.4), we will denote all probabilities and expectations related to it as P_{H, x_0} and E_{H, x_0}, accordingly. That is, we are allowing the system to start in an arbitrary state x_0, when all the on/off processes are in off periods, with H describing the joint distribution of the remainders of the initial off periods. Let

$$\tau(L) = \inf\{t \geq 0 \colon X^{(k)}(t) \geq L\}. \tag{4.8}$$

Our proof of the following theorem, which is the main result of this section, requires an assumption that the off times are not "very long." Specifically, we assume there exist a proper distribution function H^* on $[0, \infty)$ and a $1 < p < \alpha$ with

$$\int_0^{\infty} x^p \, H^*(dx) < \infty, \qquad \frac{\bar{F}_{\text{off}}(x+y)}{\bar{F}_{\text{off}}(y)} \leq 1 - H^*(x) \tag{4.9}$$

for every $x \geq 0$ and $y \geq 0$ such that $\bar{F}_{\text{off}}(y) > 0$. The class of off distributions satisfying (4.9) contains, for example, all Gamma distributions, distributions with compact support and many others.

THEOREM 4.1. *Let*

$$\bar{F}_{\text{on}}(x) = x^{-\alpha} L(x), \qquad \alpha > 1, \ x \to \infty,$$

and assume that the off distribution satisfies assumption (4.9). If the service rate r satisfies (4.5) and $r < 1$, then for any H and x_0 we have

$$\lim_{L \to \infty} \bar{F}_{\text{on}}\left(\frac{L}{1 - r + (k-1)\mu_{\text{on}}/\mu}\right) E_{H, x_0} \tau(L) = \frac{1}{k}\mu. \tag{4.10}$$

REMARKS. (i) Assumption (4.9) is not used in the proof of the lower bound in (4.10).

(ii) Without assumption (4.9) one still has an upper bound with the same order of magnitude:

$$\lim_{L \to \infty} \bar{F}_{\text{on}}\left(\frac{L}{1-r}\right) E_{H, x_0} \tau(L) \leq \frac{1}{k}\mu,$$

and it is likely that Theorem 4.1 is valid even without assumption (4.9).

An argument identical to that of Theorem 4.1 immediately proves the corresponding result for the corresponding $GI/G/1$ queue. Let $\{Y_i^{(j)},\ i \geq 1\}$, $j = 1, \ldots, k$, be iid copies of the sequence of interarrival times $\{Y_i, i \geq 1\}$, and let $\{B_i^{(j)}, i \geq 1\}$, $j = 1, \ldots, k$, be iid copies of the sequence of offered work $\{B_i, i \geq 1\}$, so that at time $S_n^{(j)} = Y_1^{(j)} + \cdots + Y_n^{(j)}$ an amount of work $B_n^{(j)}$ is brought into the system on the jth channel ($n \geq 1, 1 \leq j \leq k$). Let r be the service rate. Note that the following theorem does not require the assumption $r < 1$.

THEOREM 4.2. *Let the distribution F_{on} of B_i satisfy*

$$\bar{F}_{\text{on}}(x) = x^{-\alpha} L(x), \qquad \alpha > 1,\ x \to \infty,$$

and assume that for some $p > 1$ we have $EY_1^p < \infty$. If

$$k \frac{\mu_{\text{on}}}{\mu_{\text{off}}} < r, \tag{4.11}$$

then

$$\lim_{L \to \infty} \bar{F}_{\text{on}}(L) E\tau(L) = \frac{1}{k} \mu_{\text{off}}, \tag{4.12}$$

where $\tau(L)$ is the first time the amount of work in the system reaches the level L.

In particular, the expected time to reach a high level L in (2.12) *does not depend on the service rate r*. Note that the result of Theorem 4.2 remains true if one initializes in an arbitrary way the state of the system.

Our results show the benefits of pooling system resources. Think of k iid $GI/G/1$ queues, with holding capacity L each, and service rates r each. The queue number j is driven by the sequences $\{Y_i^{(j)}, i \geq 1\}$ and $\{B_i^{(j)}, i \geq 1\}$ as above, $j = 1, \ldots, k$. Pooling system resources means that we put together the service resources to create a "superserver" with service rate kr, and we feed this superserver by a combined stream of the k input processes, as in Theorem 4.2. The holding capacity of the new system is taken to be kL, again, as the result of pooling the resources. Consider a generic stream of "customers" or work (i.e., one of the k original streams of work). Imagine that when the holding capacity of the system serving these "customers" is reached, the system is blocked for a time to any future arrivals. Under the "k separate servers" scenario, the expected time until the serving system is blocked is $E\tau(L)$, while when the system resources are pooled, this expected time is $E\tau_{kL}$. By Theorem 4.2, the asymptotic ratio R of the two expected times is

$$R = \lim_{L \to \infty} \frac{k \bar{F}_{\text{on}}(kL)}{\bar{F}_{\text{on}}(L)} = \frac{1}{k^{\alpha - 1}} < 1,$$

which is the expected benefit of pooling the resources. However, this benefit becomes less pronounced if α is close to 1.

We are ready now for the proofs.

PROOF OF THEOREM 2.1. Choose a $\beta > 0$ small enough so that

$$1 - r \geq \beta(k-1)\frac{\mu_{\text{on}}}{\mu}. \tag{4.13}$$

Call an on period *long* if its length exceeds

$$L^* := L(1+\beta)/(1-r+(k-1)\mu_{\text{on}}/\mu).$$

Define

$$\tau^*(L) = \inf\{t \geq 0\text{: one of the } k \text{ on/off processes begins at time } t \text{ a ``long'' on period and during this time the other } (k-1) \text{ sources bring in at least } L^*(1-\beta^2)/(1+\beta) \text{ units of work}\}. \tag{4.14}$$

Observe that at time

$$\tau^*(L) + L^*$$

the state of the system is at least L, implying that

$$E_{H,x_0}\tau(L) \leq E_{H,x_0}\tau^*(L) + L^*.$$

Therefore,

$$\begin{aligned} &\limsup_{L\to\infty} \bar{F}_{\text{on}}\left(\frac{L}{1-r+(k-1)\mu_{\text{on}}/\mu}\right) E_{H,x_0}\tau(L) \\ &\quad \leq \limsup_{L\to\infty} \bar{F}_{\text{on}}\left(\frac{L}{1-r+(k-1)\mu_{\text{on}}/\mu}\right) E_{H,x_0}\tau^*(L). \end{aligned} \tag{4.15}$$

For $j = 1, \ldots, k$ and $m \geq 1$, let $Z_j^{(m)}$ be lengths of successive "long" on periods in the jth input and $\tau_m^{*,j}(L)$ is the time the mth "long" period of the jth process commences. Define

$$\tau_m^*(L) := \bigwedge_{j=1}^{k} \tau_m^{*,j}(L) \tag{4.16}$$

to be the earliest time any line commences its mth "long" period.

Observe that

$$\tau^*(L) \leq \inf\{\tau_m^*(L)\text{: during the transmission beginning at } \tau_m^*(L) \text{ the other } (k-1) \text{ sources bring at least } L^*((1-\beta^2)/(1+\beta))(k-1)(\mu_{\text{on}}/\mu) \text{ units of work}\}. \tag{4.17}$$

Therefore,

$$E_{H,x_0}\tau^*(L) \leq \sum_{m=1}^{\infty} E_{H,x_0}[\tau_m^*(L)1(A_m)], \tag{4.18}$$

with

$$A_m = \bigcap_{i=1}^{m-1} \{\text{during the transmission beginning at } \tau_i^*(L), \text{ the other } (k-1) \text{ sources bring less than } L^*((1-\beta^2)/(1+\beta))(k-1)(\mu_{\text{on}}/\mu) \text{ units of work, while during the transmission beginning at } \tau_m^*(L), \text{ the other } (k-1) \text{ sources bring at least } L^*((1-\beta^2)/(1+\beta))(k-1)(\mu_{\text{on}}/\mu) \text{ units of work}\}.$$

We remark that the asymptotics of $E_{H,x_0}\tau_m^*(L)$ as $L \to \infty$ do not change if we shorten the initial off periods by the same random amount with a finite mean. Therefore, we may assume, for now, that, under P_{H,x_0}, we have $\bigwedge_{j=1,\ldots,k} D_j^{(0)} = 0$ with probability 1. Such a change will also come in handy for the proof of the lower bound in (4.10).

Observe that by assumption (4.9) we have, for all $m \geq 1$, that, with H^* as given by (4.9),

$$P(A_m) \leq \big[kP(\text{during } L^* \text{ units of time the total amount of work brought by } (k-1) \text{ sources, each one starting with a special off period distributed according to } H^* \text{ is less than } L^*((1-\beta^2)/(1+\beta))(k-1)\mu_{\text{on}}/\mu)\big]^{m-1} := (kp_L)^{m-1}. \tag{4.19}$$

We have

$$p_L \leq (k-1)P\big(\text{during } L^* \text{ units of time the total amount of work brought by a single source starting with a special off period distributed according to } H^* \text{ is less than } L^*((1-\beta^2)/(1+\beta))\mu_{\text{on}}/\mu\big) \tag{4.20}$$

$$= \varepsilon(L) \to 0$$

as $L \to \infty$, by the law of large numbers.

Now let U be a generic random variable with the law of $\tau_1^{*,1}(L) + \hat{X}$, where the law of $\tau_1^{*,1}(L)$ is taken in the case when the source starts with an ordinary off period, and $\hat{X}$ has the conditional law of a generic on period X given $X > L^*$. Think of U as the time the first "long" period ends in process 1. Let $1 < p < \alpha$ be the number from condition (4.9). Observe that there is a finite constant c_1 such that for all $L \geq 1$ we have $E\hat{X}^p \leq c_1 L^p$. Furthermore, we claim that there is another positive constant c_2 such that for all $L \geq 1$ we have

$$E\tau_1^{*,1}(L)^p \leq c_2(\bar{F}_{\text{on}}(L))^{-p}. \tag{4.21}$$

We will check (4.21) later [see (4.33)]. Clearly, the assumption that $\bigwedge_{j=1,\ldots,k} D_j^{(0)} = 0$ implies that

$$\tau_m^*(L) \leq_{\text{st}} \sum_{i=1}^{m} U_i,$$

where $U_1, U_2, \ldots$ are iid copies of U above. Therefore, we conclude that

$$E_{H,x_0}(\tau_m^*(L))^p \le c_3 m^p (\bar{F}_{\text{on}}(L))^{-p} \tag{4.22}$$

for all $L \ge 1$, where c_3 is a finite positive constant. Letting q be the conjugate of p ($1/p + 1/q = 1$), we conclude by (4.20), (4.21) and (4.22) that

$$\begin{aligned}\sum_{m=2}^{\infty} E_{H,x_0}[\tau_m^*(L)1(A_m)] &\le \sum_{m=2}^{\infty} [E_{H,x_0}(\tau_m^*(L))^p]^{1/p} (P(A_m))^{1/q} \\ &\le c_4 (\bar{F}_{\text{on}}(L))^{-1} \sum_{m=2}^{\infty} m(k(k-1)\varepsilon(L))^{(m-1)/q} \\ &= o((\bar{F}_{\text{on}}(L))^{-1})\end{aligned}$$

as $L \to \infty$, with c_4 being once again a finite positive constant. It follows from (4.21) then that

$$\begin{aligned}&\limsup_{L\to\infty} \bar{F}_{\text{on}}\left(\frac{L}{1-r+(k-1)\mu_{\text{on}}/\mu}\right) E_{H,x_0}\tau^*(L) \\ &\quad \le \limsup_{L\to\infty} \bar{F}_{\text{on}}\left(\frac{L}{1-r+(k-1)\mu_{\text{on}}/\mu}\right) E_{H,x_0}\tau_1^*(L),\end{aligned} \tag{4.23}$$

and we proceed now to evaluate the latter limit.

Observe that $\tau_1^{*,1}(L), \ldots, \tau_1^{*,k}(L)$ are, conditionally on $(D_1^{(0)}, \ldots, D_k^{(0)})$, independent, and that

$$\tau_1^{*,j}(L) =_d D_j^{(0)} + \sum_{i=1}^{G_L^{(j)}} (X_i^{(*,j)} + Y_i^{(j)}),$$

where $G_L^{(j)}$ is a geometric random variable with parameter $\bar{F}_{\text{on}}(L^*)$, independent of two independent iid sequences, $\{X_i^{(*,j)}, i = 1, 2, \ldots\}$ and $\{Y_i^{(j)}, i = 1, 2, \ldots\}$, where the latter sequence has, as usual, the F_{off} distribution, and the former has the distribution

$$\begin{aligned}P(X_i^{(*,j)} \in A) &= P(X_1^{(j)} \in A | X_1^{(j)} \le L^*) \\ &= \frac{1}{F_{\text{on}}(L^*)} \int_0^{L^*} 1(x \in A)\, F_{\text{on}}(dx).\end{aligned}$$

Everything is also independent of the delay random variable $D_0^{(j)}$. In particular, $(X_1^{(j)} | X_1^{(j)} \le L) \le_{\text{st}} X_1^{(j)}$, and hence

$$\tau_1^{*,j}(L) \le_{\text{st}} S^{(j)}_{G_L^{(j)}}, \tag{4.24}$$

where all the random variables appearing in the right-hand side of (4.24) are independent. We conclude that

$$\tau_1^*(L) \le_{\text{st}} \bigvee_{1\le j\le k} D_0^{(j)} + \sum_{i=1}^{\wedge_{1\le j\le k} G_L^{(j)}} (X_i^{(1)} + Y_i^{(1)}),$$

and so

$$E_{H,x_0}\tau_1^*(L) \le E_{H,x_0} \bigvee_{1\le j\le k} D_0^{(j)} + \mu E\Big(\bigwedge_{1\le j\le k} G_L^{(j)} \Big).$$

Since $\wedge_{1\le j\le k} G_L^{(j)}$ is, once again, a geometric random variable with parameter $1 - F_{\text{on}}^k(L^*)$, we obtain immediately that

$$E_{H,x_0}\tau_1^*(L) \le E_{H,x_0} \bigvee_{1\le j\le k} D_0^{(j)} + \mu\Big(\frac{1}{1 - F_{\text{on}}^k(L^*)} - 1 \Big),$$

which implies that

$$\limsup_{L\to\infty} \bar{F}_{\text{on}}\Big(\frac{L}{1 - r + (k-1)\mu_{\text{on}}/\mu} \Big) E_{H,x_0}\tau(L)$$
$$\le (1+\beta)^\alpha \limsup_{L\to\infty} \bar{F}_{\text{on}}(L^*) E_{H,x_0}\tau^*(L) \le (1+\beta)^\alpha \frac{1}{k}\mu.$$

Since we can take β as close to zero as we wish, we obtain

$$\limsup_{L\to\infty} \bar{F}_{\text{on}}\Big(\frac{L}{1 - r + (k-1)\mu_{\text{on}}/\mu} \Big) E_{H,x_0}\tau(L) \le \frac{1}{k}\mu. \tag{4.25}$$

For a lower bound, we start with observing that we may, once again, assume that $\wedge_{1\le j\le k} D_j^{(0)} = 0$, P_{H,x_0}-almost surely. Indeed, shortening all the initial off periods by the same amount can only make the level crossing time smaller. To simplify the notation we will drop the subscript 1 in the definition of $\tau^*(L)$, let $\beta = 0$ and let $\tau^*(L)$ be the first time an on period of length at least

$$L_1 := \frac{L}{1 - r + (k-1)\mu_{\text{on}}/\mu} \tag{4.26}$$

begins in any line. So when $H = \delta_{(0,\dots,0)}$ and $x_0 = 0$

$$\tau^*(L) = \bigwedge_{i=1}^k \inf\{S_n^{(i)} \colon X_{n+1}^{(i)} > L_1\}. \tag{4.27}$$

[The random variable defined by (4.14) will not be used again in the proof.]

Now, take an $0 < \varepsilon < 1$, and observe that

$$\tau(L) \ge \tau^*(L(1-\varepsilon)) 1_{[\tau(L)\ge\tau^*(L(1-\varepsilon))]},$$

and so

$$E_{H,x_0}\tau(L) \ge E_{H,x_0}\tau^*(L(1-\varepsilon)) - E_{H,x_0}\big(\tau^*(L(1-\varepsilon)) 1_{[\tau(L)<\tau^*(L(1-\varepsilon))]}\big). \tag{4.28}$$

The main part of the proof of a lower bound on the expected crossing time is a proof of the fact that

$$\lim_{L\to\infty} \bar{F}_{\text{on}}(L) E_{H,x_0}(\tau^*(L(1-\varepsilon)) 1_{[\tau(L)<\tau^*(L(1-\varepsilon))]}) = 0. \tag{4.29}$$

Suppose that (4.29) has been proved. If we can establish that

$$\liminf_{L\to\infty} \bar{F}_{\mathrm{on}}(L_1)E_{H,x_0}\tau^*(L) \geq \frac{1}{k}\mu, \tag{4.30}$$

then the regular variation of $\bar{F}_{\mathrm{on}}(L)$ will provide the required counterpart to (4.25), and so prove the theorem. To prove (4.30) we may, of course, assume that $H = \delta_{(0,\ldots,0)}$ and $x_0 = 0$. We therefore use P and E without any subscripts. We remark at this point this assumption is made only for the purpose of proving (4.28) and will be removed once the latter has been proved. However, at certain later stages of the proof of the theorem we will find it useful (and possible) to reimpose this assumption.

Define the geometrically distributed random variable

$$\nu_{L_1}^{(i)} := \inf\{n\colon X_{n+1}^{(i)} > L_1\} \tag{4.31}$$

so that, as $L \to \infty$,

$$\frac{\nu_{L_1}^{(i)}}{E(\nu_{L_1}^{(i)})} \Rightarrow E^{(i)},$$

a unit exponential random variable, and

$$E(\nu_{L_1}^{(i)}) = \frac{F_{\mathrm{on}}(L_1)}{1-F_{\mathrm{on}}(L_1)} \sim \frac{1}{1-F_{\mathrm{on}}(L_1)}.$$

So

$$\inf\{S_n^{(i)}\colon X_{n+1}^{(i)} > L_1\} = S_{\nu_{L_1}^{(i)}}^{(i)}.$$

We have

$$S_{\nu_{L_1}^{(i)}}^{(i)} \sim \mu\nu_{L_1}^{(i)}$$

and so

$$(1-F_{\mathrm{on}}(L_1))S_{\nu_{L_1}^{(i)}}^{(i)} \Rightarrow \mu E^{(i)},$$

and therefore by Fatou's lemma,

$$\begin{aligned}\liminf_{L\to\infty} E((1-F_{\mathrm{on}}(L_1))\tau^*(L)) &\geq E\Big(\liminf_{L\to\infty}(1-F_{\mathrm{on}}(L_1))\tau^*(L)\Big)\\ &= E\Big(\liminf_{L\to\infty}\bigwedge_{i=1}^{k}(1-F_{\mathrm{on}}(L_1))S_{\nu_{L_1}^{(i)}}^{(i)}\Big)\\ &= E\Big(\mu\bigwedge_{i=1}^{k}E^{(i)}\Big) = \mu/k.\end{aligned}$$

So it remains to prove (4.29).

Let $p > 1$ be the number from (4.21). We notice that, by assumption (4.9), $EY_1^p < \infty$. By Hölder's inequality,

$$\begin{aligned} &E_{H,x_0}\big(\tau^*(L(1-\varepsilon))1(\tau(L) < \tau^*(L(1-\varepsilon)))\big) \\ &\quad \le \big(E_{H,x_0}(\tau^*(L(1-\varepsilon)))^p\big)^{1/p}\big(P_{H,x_0}(\tau(L) < \tau^*(L(1-\varepsilon))\big)^{1/q}, \end{aligned} \tag{4.32}$$

where $p^{-1} + q^{-1} = 1$. Let j_0 be the index where $\min\{D_1^{(0)}, \dots, D_k^{(0)}\}$ is achieved (with ties broken in, say, lexicographical manner), and recall that $D_{j_0}^{(0)} = 0$, P_{H,x_0}-almost surely. We have by the triangle inequality

$$\begin{aligned} \big(E_{H,x_0}(\tau^*(L(1-\varepsilon)))^p\big)^{1/p} &\le \big(E_{H,x_0}(\tau^*(L))^p\big)^{1/p} \le \Big(E_{H,x_0}\Big(S^{(j_0)}_{\nu^{(j_0)}_{L_1}}\Big)^p\Big)^{1/p} \\ &= \left(E\left(\sum_{i=1}^{\infty}(X_i^{(1)} + Y_i^{(1)})1_{[i\le \nu^{(1)}_{L_1}]}\right)^p\right)^{1/p} \\ &\le \sum_{i=1}^{\infty}\Big(E\Big(\Big(X_i^{(1)} + Y_i^{(1)}\Big)1_{[i\le \nu^{(1)}_{L_1}]}\Big)^p\Big)^{1/p} \\ &= \Big(E\Big(X_1^{(1)} + Y_1^{(1)}\Big)^p\Big)^{1/p}\sum_{i=1}^{\infty} P(\nu^{(1)}_{L_1} \ge i)^{1/p} \\ &\le C\bar{F}_{\mathrm{on}}(L)^{-1}, \end{aligned} \tag{4.33}$$

where C is a finite positive constant. It follows from (4.32) and (4.33) that we will prove (4.29) by showing that

$$\lim_{L\to\infty} P_{H,x_0}(\tau(L) < \tau^*(L(1-\varepsilon))) = 0. \tag{4.34}$$

Let us prove first that for every

$$N > 2k\frac{\alpha}{\alpha - 1} \tag{4.35}$$

we have

$$\lim_{L\to\infty} P_{H,x_0}(\tau(L) < \tau^*(L/N)) = 0. \tag{4.36}$$

To this end, let us "unpool" the system. That is, imagine k separate stable fluid queues defined by

$$dX_j(t) = Z_j(t)\,dt - \frac{1}{k}r(X_j(t))\,dt, \qquad t \ge 0, \tag{4.37}$$

where $\{Z_j(t), t \ge 0\}$ is given by (4.7), and $X_j(0) = (1/k)x_0$, $j = 1, \dots, k$. The k processes $\{X_j(t), t \ge 0\}$, $j = 1, \dots, k$, are, conditionally on the initial delay $(D_1^{(0)}, \dots, D_k^{(0)})$, independent. Let $Y^{(k)}(t) = X_1(t) + \cdots + X_k(t)$, $t \ge 0$. The two processes $\{X^{(k)}(t), t \ge 0\}$ and $\{Y^{(k)}(t), t \ge 0\}$ describe the states of two queuing systems. Obviously, $X^{(k)}(0) = Y^{(k)}(0)$, the two systems have identical inflow streams of work, while the outflow of work from $X^{(k)}(\cdot)$ when

the system is not empty is always at rate r, and the outflow rate from $Y^{(k)}(\cdot)$ does not exceed r. Therefore, for every ω,

$$X^{(k)}(t) \le Y^{(k)}(t), \qquad t \ge 0. \tag{4.38}$$

Note that (4.38) is just an expression of the benefit of pooling system resources. Define

$$\tau^{(Y)}(L) = \inf\{t \ge 0\colon Y^{(k)}(t) \ge L\}. \tag{4.39}$$

Then (4.38) implies that $\tau^{(Y)}(L) \le \tau(L)$, so that

$$\{\omega\colon \tau(L) < \tau^*(L/N)\} \subseteq \{\omega\colon \tau^{(Y)}(L) < \tau^*(L/N)\},$$

and, therefore,

$$P_{H,x_0}(\tau(L) < \tau^*(L/N)) \le P_{H,x_0}(\tau^{(Y)}(L) < \tau^*(L/N)). \tag{4.40}$$

Now, define

$$\tau^{(j)}(L) = \inf\{t \ge 0\colon X_j(t) \ge L\}, \qquad 1 \le j \le k.$$

Then

$$\begin{aligned} &P_{H,x_0}\big(\tau^{(Y)}(L) < \tau^*(L/N)\big) \\ &\quad \le P_{H,x_0}\big(\tau^{(j)}(L/k) < \tau^*(L/N) \text{ for some } j = 1,\dots,k\big) \\ &\quad \le \sum_{j=1}^{k} P_{H,x_0}(\tau^{(j)}(L/k) < \tau^*(L/N)) \qquad (4.41) \\ &\quad \le \sum_{j=1}^{k} P_{H,x_0}(\tau^{(j)}(L/k) < \tau^{*,j}(L/N)), \end{aligned}$$

where we recall $\tau^{*,j}$ is the time when the first long on period of length at least L_1/N begins in line j. Therefore, (4.36) will follow from (4.40) and (4.41) once we prove that for every

$$M > 2\frac{\alpha}{\alpha - 1} \tag{4.42}$$

we have

$$\lim_{L\to\infty} P_{H,x_0}(\tau^{(j)}(L) < \tau^{*,j}(L/M)) = 0, \tag{4.43}$$

for $j = 1,\dots,k$.

We will prove (4.43) for $j = 1$. Let $T_0 = \inf\{t > D_1^{(0)}\colon X_1(t) = 0\}$. Write

$$\begin{aligned} P_{H,x_0}(\tau^{(1)}(L) < \tau^{*,1}(L/M)) &= P_{H,x_0}(T_0 \le \tau^{(1)}(L) < \tau^{*,1}(L/M)) \\ &\quad + P_{H,x_0}(\tau^{(1)}(L) < \tau^{*,1}(L/M) \wedge T_0). \end{aligned}$$

Since, for all $L > 2x_0$,

$$P_{H,x_0}(\tau^{(1)}(L) < \tau^{*,1}(L/M) \wedge T_0) \le P_{H,x_0}(\tau^{(1)}(L) < T_0) \to 0$$

as $L \to \infty$ because of the rate condition (4.5) (or recall Corollary 2.2), (4.43) will follow if we prove that

$$\lim_{L\to\infty} P_{H,x_0}(T_0 \le \tau^{(1)}(L) < \tau^{*,1}(L/M)) = 0. \tag{4.44}$$

Clearly, at time T_0 the system is in an off period. Denote by $\tilde{H}$ the law (under P_{H,x_0}) of the remainder of this off period after time T_0. Since we have

$$P_{H,x_0}(T_0 \le \tau^{(1)}(L) < \tau^{*,1}(L/M)) \le P_{\tilde{H},0}(\tau^{(1)}(L) < \tau^{*,1}(L/M)),$$

(4.44) will follow once we prove (4.43) with $x_0 = 0$, and so (4.43) in its generality will follow if we prove it only for $x_0 = 0$. Assume, therefore, that $x_0 = 0$.

For $K_1 \ge 0$ and $K_2 > 0$ let $\{X_1^{(K_1,K_2)}(t), t \ge 0\}$ denote the process given by (4.37) (with $j = 1$) when $D_1^{(0)}$ is replaced by $D_1^{(0)} \wedge K_1$, and each $Y_i^{(1)}$ is replaced by $Y_i^{(1)} \wedge K_2$. Let $\tau^{(1)}(L; K_1, K_2)$ and $\tau^{*,1}(L; K_1, K_2)$ be the random times analogous to $\tau^{(1)}(L)$ and $\tau^{*,1}(L)$ correspondingly, defined with respect to the process $\{X_1^{(K_1,K_2)}(t), t \ge 0\}$. Observe that the event

$$A_{K_1,K_2} = \{\omega\colon \tau^{(1)}(L; K_1, K_2) < \tau^{*,1}_{L/M}(K_1, K_2)\}$$

increases when K_1 and K_2 decrease. It is enough, therefore, to prove (4.43) in the case when $K_1 = 0$, and K_2 is any finite positive number such that

$$\frac{\mu_{\text{on}}}{\mu_{\text{on}} + E(Y_1^{(1)} \wedge K_2)} < \frac{r}{k}.$$

In other words, we will prove (4.43) in the case when $D_1^{(0)} = 0$, the *off* times are bounded (by K_2), and $x_0 = 0$. We will, therefore, use once again P and E without any subscripts.

We define three events $A_i(L)$, $i = 1, 2, 3$, corresponding to the following three possibilities:

1. $\tau^{*,1}(L/M) < \tau^{(1)}(L) \wedge T_0$; that is, the process $\{X_1(t), t \ge 0\}$ begins an on interval of length at least L_1/M before reaching either level L or returning to 0;
2. $\tau^{(1)}(L) < \tau^{*,1}(L/M) \wedge T_0$; in other words, the process $\{X_1(t), t \ge 0\}$ reaches level L before starting an on interval of length at least L_1/M and before returning to 0;
3. $T_0 < \tau^{(1)}(L) \wedge \tau^{*,1}(L/M)$; in other words, the process $\{X_1(t), t \ge 0\}$ returns to 0 before reaching level L, and without initiating an on interval of length at least L_1/M.

Clearly,

$$\begin{aligned} P(\tau^{(1)}(L) < \tau^{*,1}(L/M)) = {} & P\big(\{\tau^{(1)}(L) < \tau^{*,1}(L/M)\} \cap A_2(L)\big) \\ & + P\big(\{\tau^{(1)}(L) < \tau^{*,1}(L/M)\} \cap A_3(L)\big). \end{aligned}$$

However, by regeneration,

$$P\big(\{\tau^{(1)}(L) < \tau^{*,1}(L/M)\} \cap A_3(L)\big) = P(A_3(L))P(\tau^{(1)}(L) < \tau^{*,1}(L/M)).$$

Therefore,

$$P(\tau^{(1)}(L) < \tau^{*,1}(L/M)) = \frac{P(\{\tau^{(1)}(L) < \tau^{*,1}(L/M)\} \cap A_2(L))}{1 - P(A_3(L))}. \tag{4.45}$$

Let $\xi_i = (1 - r/k)X_i^{(1)} - (r/k)Y_i^{(1)}$, $i = 1, 2, \ldots$. Taking into account that the off times are bounded, we can conclude that

$$P(\{\tau^{(1)}(L) < \tau^{*,1}(L/M)\} \cap A_2(L)) \le \left(P\left(\bigvee_{n=0}^{\infty} S_n^{(\xi)} > L_1/M\right)\right)^{[M/2]},$$

and $P(\bigvee_{n=0}^{\infty} S_n^{(\xi)} > L)$ is regularly varying in L with index $-(\alpha-1)$. Moreover,

$$1 - P(A_3(L)) \ge P(A_1(L)) \ge P(X_1^{(1)} > L_1/M) = \bar{F}_{\mathrm{on}}(L_1/M).$$

We conclude by (4.45) and (4.42) that

$$\limsup_{L\to\infty} P(\tau^{(1)}(L) < \tau^{*,1}(L/M)) \le \limsup_{L\to\infty} \frac{\left(P(\bigvee_{n=0}^{\infty} S_n^{(\xi)} > L_1/M)\right)^{[M/2]}}{\bar{F}_{\mathrm{on}}(L_1/M)} = 0.$$

This proves (4.43) and so (4.36) is proven as well. We observe at this point that the above argument that allowed us to assume the initial delays being equal to 0 shows that we have also proved that for every N satisfying (4.35),

$$\lim_{L\to\infty} \sup_{H} P_{H,0}(\tau(L) < \tau^*(L/N)) = 0. \tag{4.46}$$

The next step in the proof of (4.34) is to show that two "long" on periods are "unlikely to happen simultaneously." Formally, let

$$Q_L = \inf\{t \ge 0\text{: at time } t \text{ there are two on periods running, each of length at least } L\}. \tag{4.47}$$

We claim that there is a function $\gamma_L \to 0$ as $L \to \infty$ such that

$$\lim_{L\to\infty} P_{H,x_0}(Q_L \le \gamma_L^{-1}(\bar{F}_{\mathrm{on}}(L))^{-1}) = 0. \tag{4.48}$$

Of course, Q_L will only decrease if we assume that all delay times and off times are equal to 0, and Q_L is unaffected by x_0. We will, therefore, once again drop the subscripts from P and E and assume that all off times are equal to 0.

For $j_1, j_2 = 1, \ldots, k$, $j_1 \ne j_2$, let

$$Q_L^{(j_1,j_2)} = \inf\{t \ge 0\text{: at time } t, \text{ the processes } X_{j_1}(\cdot) \text{ and } X_{j_2}(\cdot) \text{ both have on periods running, each of length at least } L\}.$$

Then

$$Q_L = \bigwedge_{j_1 \ne j_2} Q_L^{(j_1,j_2)},$$

and so, for any $q > 0$,

$$P(Q_L \le q) \le \frac{k(k-1)}{2} P(Q_L^{(1,2)} \le q). \tag{4.49}$$

Let

$$Q_L^{(1)} = \inf\{t \geq 0\text{: at time } t,\ X_1(\cdot) \text{ begins an on period of length at least } L \text{ during which } X_2(\cdot) \text{ also begins an on period of length at least } L\} \tag{4.50}$$

with $Q_L^{(2)}$ defined similarly. Then

$$Q_L^{(1,2)} \geq Q_L^{(1)} \wedge Q_L^{(2)},$$

which means that, for any $q > 0$,

$$P\big(Q_L^{(1,2)} \leq q\big) \leq 2P\big(Q_L^{(1)} \leq q\big). \tag{4.51}$$

Let Z_k, $k = 1, 2, \ldots$, be an iid sequence, such that

$$Z_1 =_d \sum_{i=1}^{G_L} \hat{X}_i,$$

where G_L is a geometric random variable with parameter $\bar{F}_{\text{on}}(L)$, independent of an iid sequence $\hat{X}_i$, $i = 1, 2, \ldots$, with common law

$$P(\hat{X}_1 \in A) = P(X_1 \in A | X_1 \leq L) = \frac{1}{F_{\text{on}}(L)} \int_0^L 1(x \in A)\, F_{\text{on}}(dx).$$

Then Z_1 represents the first time $X_1(\cdot)$ starts an on period of length at least L. Let H_L be yet another geometric random variable, independent of the sequence Z_k, $k = 1, 2, \ldots$, this time with parameter p_L defined as follows. Let W be a random variable with distribution

$$P(W \in A) = P(X_1 \in A | X_1 > L) = \frac{1}{\bar{F}_{\text{on}}(L)} \int_L^\infty 1(x \in A)\, F_{\text{on}}(dx)$$

and independent of $X_2(\cdot)$. Recall that at time 0 the process $X_2(\cdot)$ starts an on interval. Then define

$$p_L = P\big(\inf\{S_n^{(2)}\text{: } n \geq 1,\ X_{n+1}^{(2)} \geq L\} \leq W\big) = P\Big(S_{\nu_L^{(2)}}^{(2)} \leq W\Big). \tag{4.52}$$

That is, p_L is the probability that $X^{(2)}(\cdot)$ begins an on period of length at least L during an on period of $X^{(1)}(\cdot)$, whose length is at least L. If $X^{(2)}(\cdot)$ does not start such an on period, we then have to wait till the next on period of $X^{(1)}(\cdot)$ whose length is at least L. Therefore,

$$Q_L^{(1)} \geq_{\text{st}} \sum_{n=1}^{H_L} Z_n. \tag{4.53}$$

We claim that

$$p_L \to 0 \quad \text{as } L \to \infty. \tag{4.54}$$

Indeed,

$$p_L = P\Big[S^{(2)}_{\nu^{(2)}_L} \le L\Big] + \int_L^\infty \frac{\bar F_{\mathrm{on}}(x)}{\bar F_{\mathrm{on}}(L)} P\Big[S^{(2)}_{\nu^{(2)}_L} \in dx\Big] = \mathrm{I} + \mathrm{II}. \tag{4.55}$$

Now

$$\mathrm{I} = P\big[\bar F_{\mathrm{on}}(L) S^{(2)}_{\nu^{(2)}_L} \le L \bar F_{\mathrm{on}}(L)\big] \to 0 \tag{4.56}$$

since $L\bar F_{\mathrm{on}}(L) \to 0$ and $\bar F_{\mathrm{on}}(L) S^{(2)}_{\nu^{(2)}_L} \Rightarrow \mu E^{(2)}$. Also, by Potter's bounds for ratios of regularly varying functions [see Bingham, Goldie and Teugels (1987)], we get for ε so small that $1 < \alpha - \varepsilon$ and all large L that

$$\begin{aligned} \mathrm{II} &= \int_1^\infty \frac{\bar F_{\mathrm{on}}(Lx)}{\bar F_{\mathrm{on}}(L)} P\left[\frac{S^{(2)}_{\nu^{(2)}_L}}{L} \in dx\right] \le c \int_1^\infty x^{-\alpha+\varepsilon} P\left[\frac{S^{(2)}_{\nu^{(2)}_L}}{L} \in dx\right] \\ &= E\left(\frac{S^{(2)}_{\nu^{(2)}_L}}{L}\right)^{-\alpha+\varepsilon} 1_{\left[S^{(2)}_{\nu^{(2)}_L} \ge L\right]} \to 0 \end{aligned} \tag{4.57}$$

as $L \to \infty$ since the integrand is bounded by 1 and

$$\frac{S^{(2)}_{\nu^{(2)}_L}}{L} = \frac{\bar F_{\mathrm{on}}(L) S^{(2)}_{\nu^{(2)}_L}}{L \bar F_{\mathrm{on}}(L)} \to_P \infty,$$

so $\mathrm{II} \to 0$.

Now define $\gamma_L = p_L^{1/2}$. Observe that by (4.49), (4.51) and (4.53) we have

$$P\big(Q_L \le \gamma_L^{-1} (\bar F_{\mathrm{on}}(L))^{-1}\big) \le k(k-1) P\Big(\sum_{n=1}^{H_L} Z_n \le p_L^{-1/2} (\bar F_{\mathrm{on}}(L))^{-1}\Big). \tag{4.58}$$

However,

$$\sum_{n=1}^{H_L} Z_n =_d \sum_{i=1}^{J_L} \hat X_i,$$

where J_L is a geometric random variable with parameter $p_L \bar F_{\mathrm{on}}(L)$, independent of $\{\hat X_i, i = 1, 2, \ldots\}$. Since, as $L \to \infty$,

$$(p_L \bar F_{\mathrm{on}}(L)) J_L \Rightarrow E(1),$$

where $E(1)$ is a standard exponential random variable, we immediately obtain (4.48) using (4.58) and the law of large numbers.

Let us go back now to the proof of (4.34). Observe, first of all, that for all L big enough,

$$P_{H, x_0}\big(\tau(L) < \tau^*(L(1-\varepsilon))\big) \le P_{H,0}\big(\tau(L(1-\varepsilon/2)) < \tau^*(L(1-\varepsilon))\big).$$

Therefore, it is enough to prove (4.34) for $x_0 = 0$, for the general case will follow by making ε smaller. We will, therefore, use the notation P_H and E_H, when $x_0 = 0$.

Fix any N satisfying (4.35) and big enough to make the right-hand side of (4.59) below positive, and observe that it is enough to prove (4.34) for $\varepsilon = 1/N$. Further, fix a ρ satisfying

$$k\rho \le \frac{\varepsilon}{2} - \varepsilon^2. \tag{4.59}$$

We have

$$\begin{aligned} &P_H\big(\tau(L) < \tau^*(L(1-\varepsilon))\big) \\ &\quad \le P_H\big(\tau(L) < \tau^*(L(1-\varepsilon)), \tau(L\varepsilon^2) \ge \tau^*(L\varepsilon^3), \tau(L) < Q_{\rho L}\big) \\ &\qquad + P_H\big(\tau(L\varepsilon^2) < \tau^*(L\varepsilon^3)\big) \\ &\qquad + P_H\big(\tau(L) \ge Q_{\rho L}\big). \end{aligned} \tag{4.60}$$

Observe that, by (4.36),

$$\lim_{L\to\infty} P_H(\tau(L\varepsilon^2) < \tau^*(L\varepsilon^3)) = 0.$$

Furthermore, by (4.48) and (4.25),

$$\begin{aligned} &P_H(\tau(L) \ge Q_{\rho L}) \\ &\quad \le P_H\big(Q_{\rho L} \le (\gamma_{\rho L})^{-1}(\bar{F}_{\mathrm{on}}(\rho L))^{-1}\big) + P_H\big(\tau(L) \ge (\gamma_{\rho L})^{-1}(\bar{F}_{\mathrm{on}}(\rho L))^{-1}\big) \\ &\quad \le P_H\big(Q_{\rho L} \le (\gamma_{\rho L})^{-1}(\bar{F}_{\mathrm{on}}(\rho L))^{-1}\big) + \gamma_{\rho L}\bar{F}_{\mathrm{on}}(\rho L)E_H\tau(L) \\ &\quad \to 0 \end{aligned}$$

as $L \to \infty$. Therefore, (4.34) will follow once we prove that

$$\lim_{L\to\infty} P_H\big(\tau(L) < \tau^*(L(1-\varepsilon)), \tau(L\varepsilon^2) \ge \tau^*(L\varepsilon^3), \tau(L) < Q_{\rho L}\big) = 0. \tag{4.61}$$

As a matter of fact, we will prove an even stronger statement. We will prove that

$$\lim_{L\to\infty} \sup_H P_H\big(\tau(L) < \tau^*(L(1-\varepsilon)), \tau(L\varepsilon^2) \ge \tau^*(L\varepsilon^3), \tau(L) < Q_{\rho L}\big) = 0. \tag{4.62}$$

Let

$$B(L) = \big\{\tau(L) < \tau^*(L(1-\varepsilon)), \tau(L\varepsilon^2) \ge \tau^*(L\varepsilon^3), \tau(L) < Q_{\rho L}\big\}.$$

We split this event into two events, $B_1(L)$ and $B_2(L)$, according to the following two possibilities:

1. After starting the first "wet period" in which the process reaches the level $L\varepsilon^2$, the process $X^{(k)}(\cdot)$ reaches level L before returning to 0.
2. The process $X^{(k)}(\cdot)$ returns to 0 before reaching level L.

Let us look at the event $B_1(L)$ first. Since $B_1(L) \subseteq B(L)$, at time $\tau(L\varepsilon^2)$ at most one of the k on/off processes has an on period whose length is at least ρL. Depending on whether the number of such on/off processes is 0 or 1, we split the event $B_1(L)$ into $B_{11}(L)$ and $B_{12}(L)$. Let us look, for example, at $B_{12}(L)$. The treatment of the event $B_{11}(L)$ is similar.

Since $B_{12}(L) \subseteq B(L)$, the single on period running at time $\tau(L\varepsilon^2)$ of length at least ρL, has length not exceeding $L_1(1-\varepsilon)$. Let us now modify the state of the system at time $\tau(L\varepsilon^2)$ in the following way. Bring all the work remaining in the presently running on periods as well as the subsequent work W_L brought by the other $k-1$ sources during the single on period of length at least ρL, and whose length does not exceed $L_1(1-\varepsilon)$, in one "lump" at time $\tau(L\varepsilon^2)$, and attach the time it takes to bring this work to the subsequent off periods. Obviously, this action can only make $\tau(L)$ smaller, and so it can only increase the probability of the event $B_{12}(L)$. Observe that, after this action, the state of the system does not exceed

$$L\varepsilon^2 + L(1-\varepsilon) + W_L + (k-1)\rho L.$$

Observe that $W_L \le L_1(1-\varepsilon)(1+\rho)$, which implies by (4.59) that the state of the system does not exceed $L(1-\varepsilon/2)$. We then increase, if necessary, the system state to exactly $L(1-\varepsilon/2)$ [only increasing in the process the probability of the event $B_{12}(L)$]. We conclude that, for some H_0,

$$\begin{aligned} P_H(B_{12}(L)) &\le P(W_L > L_1(1-\varepsilon)(1+\rho)) \\ &\quad + P_{H_0,\, L(1-\varepsilon/2)}\Big(\sup_{t\ge 0} V^{(k)}(t) \ge L\Big), \end{aligned} \tag{4.63}$$

where $\{V^{(k)}(t), t \ge 0\}$ is given by

$$dV^{(k)}(t) = Z^{(k)}(t)\,dt - r\,dt,$$

with $\{Z^{(k)}(t), t \ge 0\}$ given by (4.3) and (4.7). However,

$$P(W_L > L_1(1-\varepsilon)(1+\rho)) \to 0$$

as $L \to \infty$ by the law of large numbers. Moreover,

$$V^{(k)}(t) = V_1(t) + \cdots + V_k(t), \qquad t \ge 0,$$

where for $j = 1, \ldots, k$ the process $\{V_j(t), t \ge 0\}$ is defined by

$$dV_j(t) = Z_j(t)\,dt - \frac{r}{k}\,dt,$$

with $V_j(0) = L(1-\varepsilon/2)/k$, and with initial delay governed by the jth marginal law $H^{(j)}$ of H_0. We conclude immediately by (4.63) that

$$P_H(B_{12}(L)) \le k \sup_{H^{(1)}} P_{H^{(1)},\, L(1-\varepsilon/2)/k}\left(\sup_{t\ge 0} V_1(t) \ge \frac{L}{k}\right)$$

$$= kP\left(\sup_{t\ge 0} V_1(t) \ge \frac{L\varepsilon}{2k}\right),$$

where P without a subscript indicates, as usual, absence of delay and zero initial state. Because of the negative drift, we conclude that

$$\lim_{L\to\infty} \sup_H P_H(B_{12}(L)) = 0. \tag{4.64}$$

In exactly the same way one can show that

$$\lim_{L\to\infty} \sup_H P_H(B_{11}(L)) = 0. \tag{4.65}$$

Finally, we consider the event $B_2(L)$ above. Consider the two possibilities that are feasible after the process reaches 0: either after that time the state of the system reaches the level $L\varepsilon^2$ before the beginning of the first on period of length at least $L\varepsilon^3$, or not. Accordingly, by the strong Markov property,

$$P_H(B_2(L)) \le P_H\big(\tau^*(L\varepsilon^3) < \tau^*(L(1-\varepsilon))\big)$$

$$\times \left(\sup_G P_G(\tau(L\varepsilon^2) \le \tau^*(L\varepsilon^3)) + \sup_G P_G(B(L))\right).$$

Taking supremum over H, we obtain

$$\sup_H P_H(B(L)) \le \sup_H P_H(B_{11}(L)) + \sup_H P_H(B_{12}(L))$$

$$+ P\big(\tau^*(L\varepsilon^3) < \tau^*(L(1-\varepsilon))\big)$$

$$\times \left(\sup_H P_H(\tau(L\varepsilon^2) \le \tau^*(L\varepsilon^3)) + \sup_H P_H(B(L))\right),$$

which is the same as

$$\sup_H P_H(B(L)) \le \Big[\sup_H P_H(B_{11}(L)) + \sup_H P_H(B_{12}(L))$$

$$+ \sup_H P_H(\tau(L\varepsilon^2) \le \tau^*(L\varepsilon^3))\Big] \tag{4.66}$$

$$\times \big[P(\tau^*(L\varepsilon^3) = \tau^*(L(1-\varepsilon)))\big]^{-1}.$$

Now (4.62) follows from (4.66), (4.64), (4.65), (4.46) and the fact that

$$P\big(\tau^*(L\varepsilon^3) = \tau^*(L(1-\varepsilon))\big) = \frac{\bar F_{\mathrm{on}}(L(1-\varepsilon))}{\bar F_{\mathrm{on}}(L\varepsilon^3)} \to \left(\frac{\varepsilon^3}{1-\varepsilon}\right)^{\alpha} > 0$$

as $L\to\infty$ by the regular variation. This completes the proof of the theorem. □

REFERENCES

ABATE, J., CHOUDHURY, G. and WHITT, W. (1994). Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems Theory Appl.* 16 311–338.

ANICK, D., MITRA, D. and SONDHI, M. (1982). Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal* 61 1871–1894.

ASMUSSEN, S. (1987). *Applied Probability and Queues*. Wiley, New York.

ASMUSSEN, S. (1998). Subexponential asymptotics for stochastic processes: extremal behavior, stationary distributions and first passage probabilities. *Ann. Appl. Probab.* To appear.

ASMUSSEN, S. and PERRY, D. (1992). On cycle maxima, first passage problems and extreme value theory for queues. *Stochastic Models* 8 421–458.

AVRAM, F. and TAQQU, M. S. (1986). Weak convergence of moving averages with infinite variance. In *Dependence in Probability and Statistics* (E. Eberlein and M. Taqqu, eds.) 399–415. Birkhäuser, Boston.

AVRAM, F. and TAQQU, M. S. (1989). Probability bounds for M-Skorohod oscillations. *Stochastic Process. Appl.* 33 63–72.

AVRAM, F. and TAQQU, M. S. (1992). Weak convergence of sums of moving averages in the α-stable domain of attraction. *Ann. Probab.* 20 483–503.

BERGER, A. and WHITT, W. (1995). Maximum values in queueing processes. *Probab. Engrg. Inform. Sci.* 9 375–409.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

BINGHAM, N., GOLDIE, C. and TEUGELS, J. (1987). Regular variation. In *Encyclopedia of Mathematics and Its Applications* 27. Cambridge Univ. Press.

BRICHET, F., ROBERTS, J., SIMONIAN, A. and VEITCH, D. (1996). Heavy traffic analysis of a storage model with long range dependent On/Off sources. *Queueing Systems Theory Appl.* 23 197–215.

CHOUDHURY, G. and WHITT, W. (1995). Long-tail buffer-content distributions in broadband networks. Preprint, AT&T Bell Laboratories, Murray Hill, NJ.

CROVELLA, M. and BESTAVROS, A. (1995). Explaining World Wide Web traffic self-similarity. Preprint. (Available as TR-95-015 from {crovella,best}@cs.bu.edu.)

CUNHA, C., BESTAVROS, A. and CROVELLA, M. (1995). Characteristics of www client-based traces. Preprint. (Available as BU-CS-95-010 from {crovella,best}@cs.bu.edu.)

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

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* 2, 2nd ed. Wiley, New York.

HEATH, D., RESNICK, S. and SAMORODNITSKY, G. (1996). Heavy tails and long range dependence in on/off processes and associated fluid models. *Math. Oper. Res.* To appear. (Available as TR1144.ps.Z at http://www.orie.cornell.edu/trlist/trlist.html.)

IGLEHART, D. (1972). Extreme values in the GI/G/1 queue. *Ann. Math. Statist.* 43 627–635.

KELLA, O. and WHITT, W. (1992). A storage model with a two-state random environment. *Oper. Res.* 40 257–262.

LELAND, W., TAQQU, M., WILLINGER, W. and WILSON, D. (1993). On the self-similar nature of ethernet traffic. *ACM/SIGCOMM Computer Communications Review* 23 183–193.

LELAND, W., TAQQU, M., WILLINGER, W. and WILSON, D. (1994). On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking* 2 1–15.

PACHECO, A. and PRABHU, N. U. (1996). A Markovian storage model. *Ann. Appl. Probab.* 6 76–91.

PRABHU, N. U. and PACHECO, A. (1995). A storage model for data communication systems. *Queueing Systems Theory Appl.* 19 1–40.

PRATT, J. (1960). On interchanging limits and integrals. *Ann. Math. Statist.* 31 74–77.

RESNICK, S. (1986). Point processes, regular variation and weak convergence. *Adv. in Appl. Prob.* 18 66–138.

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

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

WILLINGER, W., TAQQU, M., LELAND, W. and WILSON, D. (1995). Self-similarity in high-speed packet traffic: analysis and modeling of Ethernet traffic measurements. *Statist. Sci.* 10 67–85.

WILLINGER, W., TAQQU, M., SHERMAN, R. and WILSON, D. (1995). Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. Preprint.

SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
CORNELL UNIVERSITY
RHODES HALL
ITHACA, NEW YORK 14853
E-MAIL: davidh@orie.cornell.edu
sid@orie.cornell.edu
gennady@orie.cornell.edu