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

David Heath, Sidney Resnick and Gennady Samorodnitsky

Cornell University

Abstract. We study the time it takes until a 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, and 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 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 LANs, 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 non–standard dependence structures. See for example Leland, Taqqu, Willinger and Wilson (1993, 1994); Willinger, Taqqu, Leland and Wilson (1995); Crovella and Bestavros (1995); 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 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 indicates 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 et al (1996), Kella and Whitt (1992), Choudhury and Whitt (1995).

1991 *Mathematics Subject Classification.* 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 workload, weak convergence.

David Heath, Sidney Resnick and Gennady Samorodnitsky were partially supported by NSA Grant MDA904-95-H-1036 at Cornell University. Resnick and Samorodnitsky also received support from NSF Grant DMS-94-00535 at Cornell University. Samorodnitsky also acknowledges support from the United States – Israel Binational Science Foundation (BSF) Grant 92-00074

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 $\{X_i, i = 1, 2, \dots\}$ be a sequence of iid nonnegative random variables representing *on* periods, and similarly let $\{Y_i, i = 1, 2, \dots\}$ 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 F_{off} be the common distribution of Y_i's. The workload arrives in the system at rate 1 during *on* periods (no workload arrives in the system during *off* periods). The service rate is r. That is, whenever the system is nonempty, workload is leaving the system at rate r. The state of the system at time t (its content at time t, the workload in the system at time t) is denoted by $X(t)$, and can be formally defined as follows. For a $t \geq 0$ let

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

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 workload 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 the system (1.2) to be stable. If

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

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

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

where

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

and

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

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. Secondly, the average *off* time, $\mu_{\rm off}$, 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), for in this case the expected time until a single *on* period causes the system to exceed the capacity L is asymptotic to

$$(\mu_{\rm on}+\mu_{\rm off})\exp\Big(\frac{L}{(1-r)\mu_{\rm on}}\Big),$$

which is of a larger 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 $\mu_{\rm off}$ 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 workload 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 behaviors in cases where *on* and *off* distributions have exponentially bounded tails with 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); 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), 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

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

so that

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

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, 1987; 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

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

where L is a slowly varying function. Note that the process $\{X(t), t \geq 0\}$ is regenerative (cf. Resnick, 1992; Feller, 1971; Asmussen, 1987). One set of regeneration times is

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

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 $x \to \infty$*

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

The proof of this critical result is deferred to the end of this section. Note the result depends on F_{on} and r but that F_{off} only affects the answer through the multiplicative factor $E(\bar{N})$.

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

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

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

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

associated to the queuing process $\{W_n\}$ and that it is not the downgoing ladder epoch of $\{S_n\}$ which is determining 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 $x \to \infty$*

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

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

Proof. 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

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

Thus

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

where the last step uses Proposition 2.1.

To get a reverse inequality, choose K such that

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

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

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

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

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

and applying Proposition 2.1 to the random walk $\{(1-r)S_j^{(X)} - rS_j^{(Y,K)}, j \geq 0\}$ 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. $\square$

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 non-decreasing function $U : (0,\infty) \mapsto (0,\infty)$ define the (left continuous) inverse

$$U^{\leftarrow}(x) = \inf\{s > 0 : U(s) \geq x\}, \quad x > 0.$$

Define the non-decreasing process

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

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

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

Standard inversion techniques from extreme value theory (Resnick, 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.

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

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

so that

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

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 $u \to \infty$

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

and

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

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 $\{\bar{N}_k, k \geq 1\}$ be the iterates of $\bar{N}$ so that $\bar{N}_k$ is the kth downgoing ladder epoch of the random walk $\{(1-r)S_n^{(X)} - rS_n^{(Y)}, n \geq 0\}$. 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 $\{M_i, i \geq 1\}$ 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 $\{S_{\bar{N}_k}, k \geq 0\}$, let

$$\Theta(t) = \inf\{k : S_{\bar{N}_k} \geq t\}$$

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

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

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

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

Now from Billingsley, 1968, Theorem 4.4 and composition

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

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

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

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

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

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

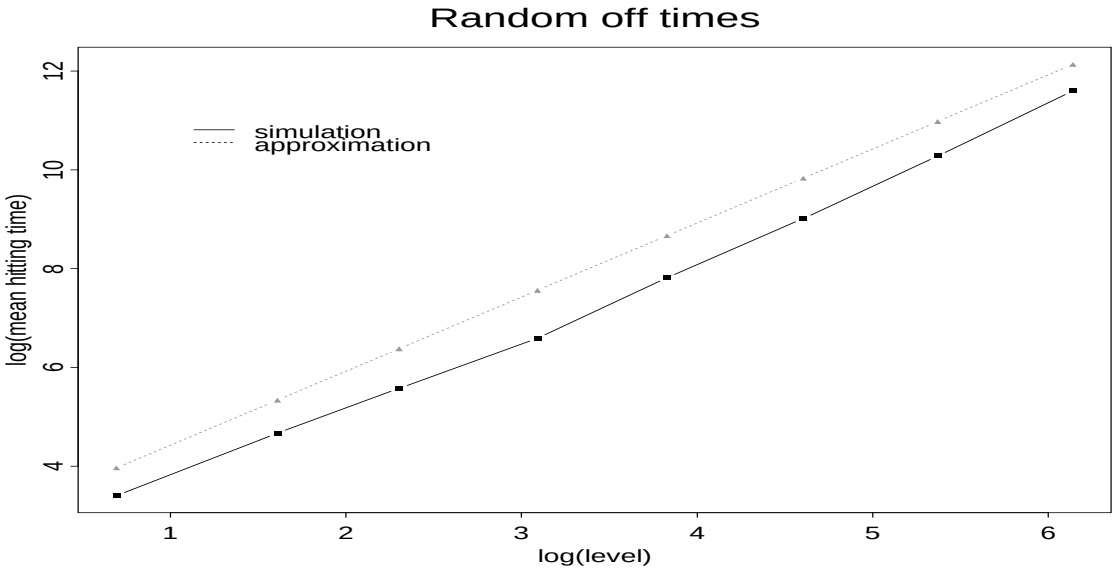

Figure 2.1. Pareto *on/off* periods, $\alpha = 1.5,\ r = .53$.

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, 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

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

Observe for $y > 0$

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

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

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

For a reverse inequality, note that

$$\tau(L) \le S_\nu$$

where

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

so that

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

However

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

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 $r = .53$. For the first simulation, F_{off} was the same Pareto and for the second simulation 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.

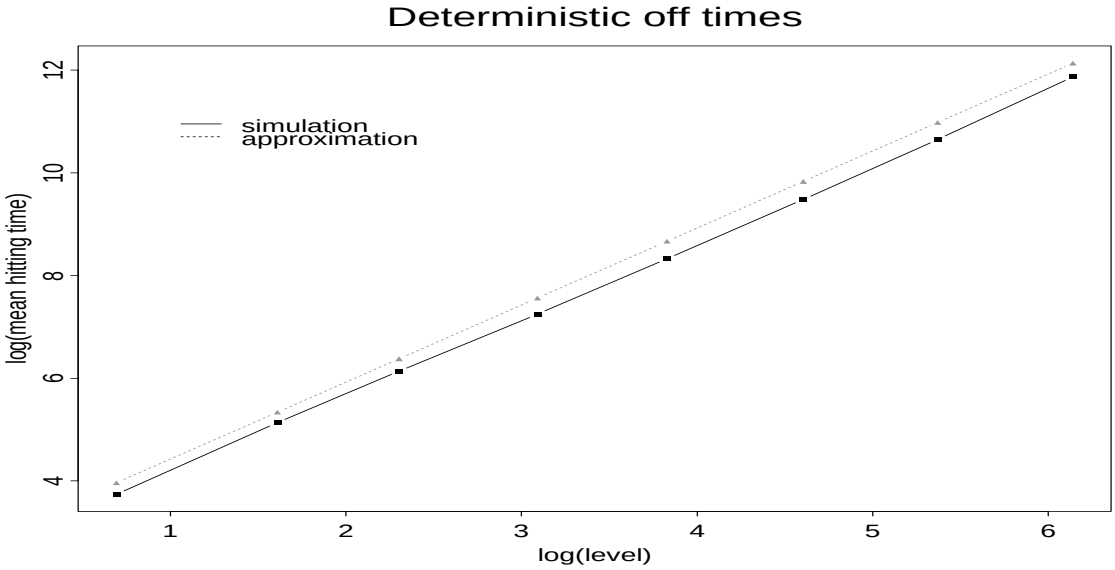

Figure 2.2. Pareto *on* period, deterministic *off* period, $\alpha = 1.5$, $r = .53$.

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 $r = .53$. 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

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

the truncation by 3 being for the purpose of keeping the data in a comfortable range. The range of the 1000 realizations was $[.896, 3]$ 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. A histogram of the data follows showing the preponderance of observations to the right of 1.

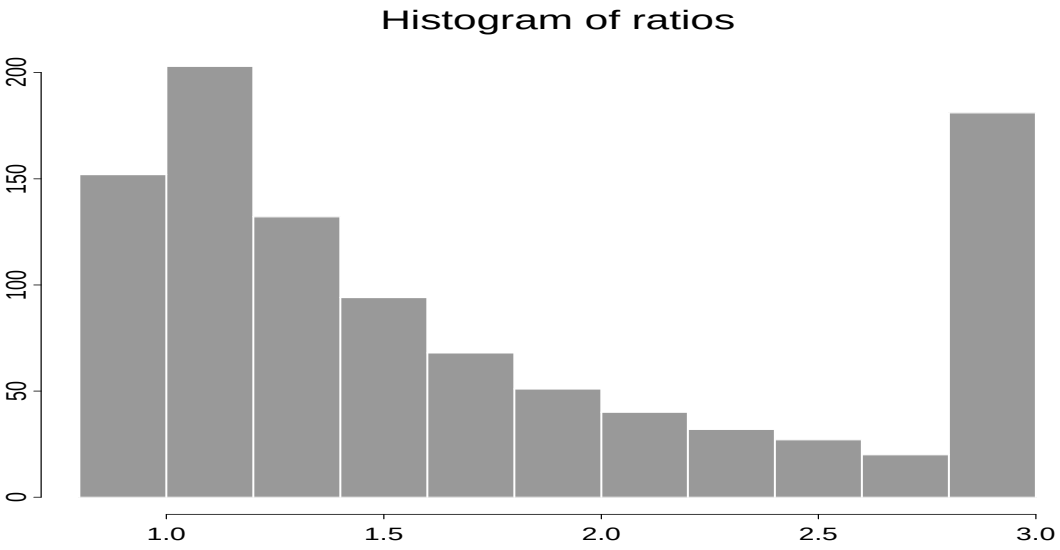

Figure 2.3. $\frac{(1-r)X_{\text{on}}}{L} \wedge 3$ for Pareto *on* period, deterministic *off* period, $\alpha = 1.5,\ r = .53$.

It remains to prove Proposition 2.1. We restate it in the following form.

Proposition 2.1′. *Suppose* $\{\xi_n, n \geq 1\}$ *are iid with* $E\xi_1 < 0$ *and for* $x > 0$

$$P[\xi_1 > x] =: \bar{F}(x) \sim x^{-\alpha} L(x), \quad \alpha > 1$$

where L *is a slowly varying function. Let*

$$\{S_n^{(\xi)}, n \geq 0\} = \{0, \sum_{i=1}^{n} \xi_i, n \geq 1\}$$

be the random walk with negative drift and step distribution F*. Suppose*

$$\bar{N} = \inf\{n > 0 : S_n^{(\xi)} \leq 0\}$$

is the downgoing ladder epoch of $\{S_n^{(\xi)}\}$*. Then as* $x \to \infty$

$$P[\bigvee_{n=0}^{\bar{N}} S_n^{(\xi)} > x] \sim E(\bar{N}) P[\xi_1 > x] = E(\bar{N}) \bar{F}(x).$$

Proof. For $x > 0$ let

$$N(x) = \inf\{n > 0 : S_n^{(\xi)} > x\}.$$

Then

$$\begin{aligned}
P[\bigvee_{n=0}^{\bar{N}} S_n^{(\xi)} > x] &= P[\bigvee_{n=0}^{\bar{N}} S_n^{(\xi)} > x, N(x) < \bar{N}] \\
&= \sum_{n=1}^{\infty} P[S_i^{(\xi)} \in [0,x],\ i = 1,\dots,n-1; S_n^{(\xi)} > x] \\
&\geq \sum_{n=1}^{\infty} P[S_i^{(\xi)} \in [0,x],\ i = 1,\dots,n-1; X_n > x] \\
&= \bar{F}(x) \sum_{n=1}^{\infty} P[S_i^{(\xi)} \in [0,x],\ i = 1,\dots,n-1] \\
&= \bar{F}(x) E(\bar{N} \wedge N(x)) \\
&\sim \bar{F}(x) E(\bar{N}),
\end{aligned}$$

as $x \to \infty$. Thus we get

$$\liminf_{x\to\infty} \frac{P[\bigvee_{n=0}^{\bar{N}} S_n^{(\xi)} > x]}{\bar{F}(x)} \geq E(\bar{N}). \tag{2.11}$$

The reverse inequality is more challenging. Observe that

$$P[\bigvee_{n=0}^{\bar{N}} S_n^{(\xi)} > x] = P[\bigvee_{n=0}^{\bar{N}\wedge N(x)} S_n^{(\xi)} > x]$$

and that

$$P[\bigvee_{n=0}^{\bar{N}} \xi_n > x] \sim P[\bigvee_{n=0}^{\bar{N}\wedge N(x)} \xi_n > x] \sim E(\bar{N})\bar{F}(x) \tag{2.12}$$

where the first equivalence follows from Resnick (1986) and second being a minor modification of the first. So we attempt to compare

$$P[\bigvee_{n=0}^{\bar{N}\wedge N(x)} S_n^{(\xi)} > x] \text{ with } P[\bigvee_{n=0}^{\bar{N}\wedge N(x)} \xi_n > x].$$

We write as an abbreviation

$$N^* = \bar{N} \wedge N(x).$$

Pick $\lambda > 0$. Then we have

$$\begin{aligned}
P[\bigvee_{n=0}^{N^*} S_n^{(\xi)} > x] - P[\bigvee_{n=1}^{N^*} \xi_n > x] & \\
&= P[\bigvee_{n=0}^{N^*} S_n^{(\xi)} > x, \bigvee_{n=1}^{N^*} \xi_n \leq x] \\
&= P[\bigvee_{n=0}^{N^*} S_n^{(\xi)} > x, \bigvee_{n=1}^{N^*} \xi_n \leq x - \lambda] + P[\bigvee_{n=0}^{N^*} S_n^{(\xi)} > x, x - \lambda < \bigvee_{n=1}^{N^*} \xi_n \leq x] \\
&= I_1(x) + I_2(x). \qquad (2.13)
\end{aligned}$$

From (2.12) we get

$$\frac{I_2(x)}{\bar{F}(x)} \leq \frac{P[\bigvee_{i=1}^{N^*} \xi_i > x-\lambda] - P[\bigvee_{i=1}^{N^*} \xi_i > x]}{\bar{F}(x)} \to 0 \tag{2.14}$$

as $x \to \infty$. So I_2 is of small order and we turn attention to I_1.

Pick and fix K such that $[K/2] > \alpha/(\alpha-1)$ and write

$$\begin{aligned} I_1(x) \leq & P[\bigvee_{i=0}^{N^*} S_i^{(\xi)} > x, \bigvee_{i=1}^{N^*} \xi_i \leq x/K] + P[\bigvee_{i=0}^{N^*} S_i^{(\xi)} > x, x/K < \bigvee_{i=1}^{N^*} \xi_i < x-\lambda] \\ = & I_{11}(x) + I_{12}(x). \end{aligned}$$

For an integer M we further decompose

$$\begin{aligned} I_{12}(x) = & P[\bigvee_{i=0}^{N^*} S_i^{(\xi)} > x, x/K < \bigvee_{i=1}^{N^*} \xi_i \leq x-\lambda, \bigvee_{i=1}^{M} \xi_i > x/K] \\ & + P[\bigvee_{i=0}^{N^*} S_i^{(\xi)} > x, \bigvee_{i=1}^{N^*} \xi_i \leq x-\lambda, \bigvee_{i=1}^{M} \xi_i \leq x/K, \bigvee_{i=M+1}^{N^*} \xi_i > x/K, N^* > M] \\ = & I_{121}(x) + I_{122}(x). \end{aligned}$$

For $x > \lambda$

$$\begin{aligned} I_{121}(x) \leq & P[\text{ There exists some } n \leq M \wedge N^* \text{ such that } \xi_n > x/K, \bigvee_{m=0}^{N^*} \sum_{\substack{j=1 \\ j\neq n}}^{m} \xi_j > \lambda] \\ \leq & P[\text{ There exists } n \leq M \text{ such that } \xi_n > x/K, \bigvee_{m=0}^{\infty} \sum_{\substack{j=1 \\ j\neq n}}^{m} \xi_j > \lambda] \\ \leq & M\bar{F}(x/K) P[\bigvee_{j=0}^{\infty} S_j^{(\xi)} > \lambda] \end{aligned}$$

and thus we conclude

$$\limsup_{x\to\infty} \frac{I_{121}(x)}{\bar{F}(x)} \leq MK^{\alpha} P[\bigvee_{j=0}^{\infty} S_j^{(\xi)} > \lambda]. \tag{2.15}$$

Furthermore for $I_{122}(x)$ we have

$$\begin{aligned} I_{122}(x) \leq & P\{\bigcup_{n\geq M} [\xi_n > x/K, S_i^{(\xi)} \in [0,x],\ i=1,\ldots,n-1]\} \\ \leq & \sum_{n=M+1}^{\infty} P[S_i^{(\xi)} \in [0,x],\ i=1,\ldots,n-1;\ \xi_n > x/K] \\ = & \bar{F}(x/K) E(N^* \wedge N(x) 1_{[N^* \wedge N(x) > M]} \end{aligned}$$

and thus

$$\limsup_{x\to\infty}\frac{I_{122}(x)}{\bar{F}(x)} \le K^\alpha E(\bar{N}1_{[\bar{N}>M]}). \tag{2.16}$$

Finally we deal with $I_{11}(x)$. Define stopping times

$$N_1(x/K) = \inf\{n \ge 0 : S_n^{(\xi)} \ge x/K\},$$
$$N_i(x/K) = \inf\{n \ge N_{i-1}(x/K) : S_n^{(\xi)} - S^{(\xi)}_{N_{i-1}(x/K)} \ge x/K\}, \quad i \ge 2.$$

Because all positive steps of the random walk must be small in the piece controlled by $I_{11}(x)$, we have

$$I_{11}(x) \le P[N_i(x/K) \le \bar{N},\ i = 1,\dots,[K/2]]$$

and by the strong Markov property, this is bounded by

$$\le P[N_i(x/K) \le \bar{N},\ i = 1,\dots,[K/2]-1]P[\bigvee_{n=0}^{\infty} S_n^{(\xi)} > x/K]$$
$$\le \cdots \le (P[\bigvee_{n=0}^{\infty} S_n^{(\xi)} > x/K])^{[K/2]}.$$

Since $P[\bigvee_{n=0}^{\infty} S_n^{(\xi)} > z]$ is regularly varying in z with index $-(\alpha-1)$ (see for example, Bingham, Goldie and Teugels, 1987, page 387) we have

$$\frac{I_{11}(x)}{\bar{F}(x)} \le \frac{(P[\bigvee_{n=0}^{\infty} S_n^{(\xi)} > x/K])^{[K/2]}}{\bar{F}(x)} \to 0$$

as $x \to \infty$.

We now must put the pieces together. We have

$$\limsup_{x\to\infty}\frac{P[\bigvee_{n=0}^{N^*} S^{(\xi)} > x] - P[\bigvee_{n=0}^{N^*}\xi_n > x]}{\bar{F}(x)} \le \limsup_{x\to\infty}\frac{I_{11}(x)+I_{121}(x)+I_{122}(x)+I_2(x)}{\bar{F}(x)}$$
$$\le K^\alpha M P[\bigvee_{n=0}^{\infty} S_n^{(\xi)} > \lambda] + K^\alpha E(\bar{N}1_{[\bar{N}>M]}).$$

Let $\lambda \to \infty$ and then $M \to \infty$ to get

$$\limsup_{x\to\infty}\frac{P[\bigvee_{n=0}^{N^*} S^{(\xi)} > x] - P[\bigvee_{n=0}^{N^*}\xi_n > x]}{\bar{F}(x)} = 0.$$

Since $P[\vee_{n=0}^{N^*}\xi_n > x] \sim E(\bar{N})\bar{F}(x)$ we get

$$\limsup_{x\to\infty}\frac{P[\bigvee_{n=0}^{N^*} S_n^{(\xi)} > x]}{\bar{F}(x)} \le E(\bar{N})$$

which is the desired reverse inequality. The proof is complete. □

3. Contrast with exponential tails. If F_{on} has an exponentially bounded tail, known results of Iglegart (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

(A1) $$\text{For some } \gamma > 0\text{: } E^{\gamma\xi_1} = 1$$

and for this value of γ we have

(A2) $$E\xi_1 e^{\gamma\xi_1} := \mu_\gamma \in (0,\infty)$$

and

(A3) $$\xi_1 \text{ has a non-lattice distribution.}$$

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 $x > 0$*

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

where

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

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

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

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

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

We then get as $u \to \infty$

$$\left(\gamma M(ut) - \log a(0)u,\ M^{\leftarrow}(\frac{x + \log a(0)u}{\gamma})/u\right) \Rightarrow (Y_0(t/(\mu E(\bar{N})), \mu E(\bar{N}) Y_0^{\leftarrow}(x))$$

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

(3.3) $$\left(\frac{a(0)}{\mu E(\bar{N})}\frac{\tau(u)}{e^{\gamma u}}\right) = \left(E(e^{\gamma(1-r)X_1})\frac{(1 - Ee^{-\gamma S_{\bar{N}}^{(\xi)}})^2}{\mu(E\bar{N})^2\gamma\mu_\gamma}\right)\frac{\tau(u)}{e^{\gamma u}} \Rightarrow E(1)$$

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

$$0 \leq \tau(s) \leq S_{\bar{N}_{V(s)}} \tag{3.4}$$

where

$$V(s) = \inf\{n : \bigvee_{i=1}^{n} M_i \geq s\}$$

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 $s \to \infty$

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

So as $s \to \infty$

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

and

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

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

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

Note how critically F_{off} enters into formulas (3.3), (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

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

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

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

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

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

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

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

With these formulas we find

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

Next observe that because of exponential tails,

$$S_{\bar{N}}^{(\xi)} \stackrel{d}{=} -r\mu_{\text{off}}E(1),$$

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

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

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

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

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

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

However from (3.5)

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

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

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

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

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

we describe the state of the system prior to reaching level s as a Markov process $\{\tilde{X}(t), t \geq 0\}$ with a 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

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

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

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

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 $t \geq 0$

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

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

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

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

with the obvious boundary conditions

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

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 $\{X_i^{(j)},\ i=1,2,\dots\}$, $j=1,\dots,k$ and $\{Y_i^{(j)},\ i=1,2,\dots\}$, $j=1,\dots,k$ be iid copies of the *on* sequence $\{X_i,\ i=1,2,\dots\}$ and the *off* sequence $\{Y_i,\ i=1,2,\dots\}$ correspondingly. We construct k iid *on/off* processes, $\{Z_j(t),\ t\geq 0\}$, $j=1,\dots,k$ as in (1.1). Sometimes we will find it convenient to work with stationary versions of $\{Z_j(t),\ t\geq 0\}$, $j=1,\dots,k$. Those exist due to the finiteness of $\mu_{\rm on}$ and $\mu_{\rm off}$, and can be constructed as follows. Fix a $j=1,\dots,k$, 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

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

and ($x>0$)

$$P[X_j^{(0)}>x]=\int_x^\infty \frac{\bar F_{\rm on}(s)}{\mu_{\rm on}}ds,\quad P[Y_j^{(0)}>x]=\int_x^\infty \frac{\bar F_{\rm off}(s)}{\mu_{\rm off}}ds.$$

Finally, $Y_0^{(j)}$ has the $F_{\rm off}$ 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

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

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 queue as above, with service rate r, fed by the k *on/off* processes. That is, the combined inflow rate in the system is given by

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

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 towards understanding the queues with more general long memory input streams and, secondly, to understand the effect of pooling the resources in the systems of the type we are considering. The natural rate condition for this system, parallel to (1.3), is

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

saying that the long-term inflow rate to the system (4.4) is less than the potential long-term outflow rate from the system. Of course, we also assume that $r < k$, to make sure that the system is non-degenerate. In the appendix we verify that, under condition (4.5), there is a stationary stochastic process $\{X^{(k)}(t),\ t \geq 0\}$ satisfying (4.4). This statement, though intuitively obvious, does require an argument due to the lack of renewal structure in the process $\{Z^{(k)}(t),\ t \geq 0\}$. We assume, as always, that the distribution $F_{\rm on} * F_{\rm off}$ 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 non-triviality 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 the 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 being "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 R_+^k, whose marginals have finite first moments. Let $(D_1^{(0)} \ldots, D_k^{(0)})$ be an H-distributed random vector, independent of the sequences $\{X_i^{(j)},\ i = 1, 2, \ldots\}$, $j = 1, \ldots, k$ and $\{Y_i^{(j)},\ i = 1, 2, \ldots\}$, $j = 1, \ldots, k$. 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 $X^{(k)}(0) = x_0 \in \{0, 1, 2, \ldots\}$ 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

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

Theorem 4.1. *Let*

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

and assume that for some $p > 1$ we have $EY_1^p < \infty$. If the service rate r satisfies (4.5), and $r < 1$, then for any H and x_0 we have

$$\lim_{L\to\infty} \bar{F}_{\rm on}\Big(\frac{L}{1-r}\Big) E_{H,x_0}\tau(L) = \lim_{L\to\infty} \bar{F}_{\rm on}\Big(\frac{L}{1-r}\Big) E_{H,x_0}\tau^*(L) = \frac{1}{k}\mu, \tag{4.9}$$

where

$$\tau^*(L) = \inf\{t \geq 0 : \textit{one of the } k \textit{ on/off processes begins at time } t \textit{ an on period of length at least} \frac{L}{1-r}\}.$$

An argument identical to that of Theorem 2.1 immediately proves the corresponding result for the corresponding $GI/G/1$ queue. Let $\{Y_i^{(j)},\ i = 1, 2, \dots\}$, $j = 1, \dots, k$ be iid copies of the sequence of interarrival times $\{Y_i,\ i = 1, 2, \dots\}$, and let $\{B_i^{(j)},\ i = 1, 2, \dots\}$, $j = 1, \dots, k$ be iid copies of the sequence of offered work $\{B_i,\ i = 1, 2, \dots\}$, so that at time $S_n^{(j)} = Y_1^{(j)} + \dots + Y_n^{(j)}$ an amount of work $B_n^{(j)}$ is brought in the system on the jth channel, $n = 0, 1, \dots$; $j = 1, \dots, k$. Let r be the service rate. Note that the following theorem does not require the assumption $r < 1$.

Theorem 2.2. *Let the distribution F_{on} of B_i satisfy*

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

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 2.2 remains true if one initializes in an arbitrary way the state of the system.

Let us look at the implications of our results on the benefits of pooling the 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 $\{Y_i^{(j)},\ i = 1, 2, \dots\}$ and $\{B_i^{(j)},\ i = 1, 2, \dots\}$ as above, $j = 1, \dots, k$. We take pooling system resources here to mean that we put together the service resources to create a "super-server" with service rate kr, and we feed this "super-server" by a combined stream of the k input processes, as in Theorem 2.2. The holding capacity of the new system is taken to be kL, again, as the result of pooling the resources. Let us look at a generic stream of "customers", or work (i.e. one of the k original streams of work). One can 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 2.2 the ratio R of the two expected times is, asymptotically for large L,

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

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. Obviously,

$$E_{H, x_0} \tau(L) \leq E_{H, x_0} \tau^*(L) + \frac{L}{1-r}. \tag{4.13}$$

If

$$\tau^{*,j}(L) = \inf\Big\{ S_n^{(j)} : \ n \geq 1, \ X_{n+1}^{(j)} \geq \frac{L}{1-r} \Big\}, \ j = 1, \ldots, k, \tag{4.14}$$

we have

$$\tau^*(L) = \min\Big(\tau^{*,1}(L), \ldots, \tau^{*,k}(L) \Big). \tag{4.15}$$

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

$$\tau^{*,j}(L) \stackrel{d}{=} D_j^{(0)} + \sum_{i=1}^{G_L^{(j)}} (X_i^{(*,j)} + Y_i^{(j)}), \tag{4.16}$$

where $G_L^{(j)}$ is a geometric random variables with parameter $\bar{F}_{\rm on}(\frac{L}{1-r})$, 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 $F_{\rm off}$ distribution, and the former has the distribution

$$P(X_i^{(*,j)} \in A) = P\Big(X_1^{(j)} \in A | X_1^{(j)} \leq L \Big) = \frac{1}{F_{\rm on}(\frac{L}{1-r})} \int_0^{L/(1-r)} 1(x \in A) \ F_{\rm on}(dx).$$

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

$$\tau^{*,j}(L) \stackrel{\rm st}{\leq} S^{(j)}_{G_L^{(j)}}, \tag{4.17}$$

where all the random variables appearing in the right hand side of (4.17) are independent. We conclude that

$$\tau^*(L) \stackrel{\rm st}{\leq} \bigvee_{1 \leq j \leq k} D_0^{(j)} + \sum_{i=1}^{\wedge_{1 \leq j \leq k} G_L^{(j)}} (X_i^{(1)} + Y_i^{(1)}),$$

and so

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

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

$$E_{H,x_0} \tau(L) \leq E_{H,x_0} \tau^*(L) + \frac{L}{1-r} \leq E_H \bigvee_{1 \leq j \leq k} D_0^{(j)} + \mu \Big(\frac{1}{1 - F_{\rm on}(\frac{L}{1-r})^k} - 1 \Big) + \frac{L}{1-r},$$

which implies that

$$\limsup_{L \to \infty} \bar{F}_{\rm on}\Big(\frac{L}{1-r} \Big) E_{H,x_0} \tau(L) \leq \limsup_{L \to \infty} \bar{F}_{\rm on}\Big(\frac{L}{1-r} \Big) E_{H,x_0} \tau^*(L) \leq \frac{1}{k} \mu. \tag{4.18}$$

For a lower bound, we start with observing that we may 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 bring the level crossing time closer. Now, take an $0 < \epsilon < 1$, and observe that

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

and so

$$E_{H,x_0}\tau(L) \ge E_{H,x_0}\tau^*(L(1-\epsilon)) - E_{H,x_0}\Big(\tau^*(L(1-\epsilon))1_{[\tau(L)<\tau^*(L(1-\epsilon))]}\Big). \tag{4.19}$$

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

$$\lim_{L\to\infty} \bar F_{\rm on}\Big(\frac{L}{1-r}\Big) E_{H,x_0}\Big(\tau^*(L(1-\epsilon))1_{[\tau(L)<\tau^*(L(1-\epsilon))]}\Big) = 0. \tag{4.20}$$

Suppose that (4.20) has been proved. If we can establish that

$$\liminf_{L\to\infty} \bar F_{\rm on}\Big(\frac{L}{1-r}\Big) E_{H,x_0}\tau^*(L) \ge \frac{1}{k}\mu, \tag{4.21}$$

then (4.19), (4.20) and the regular variation of $\bar F_{\rm on}(L)$ will provide the required counterpart to (4.18), and so prove the theorem. To prove (4.21) we may, of course, assume that $H = \delta_{(0,\dots,0)}$, 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 of (2.21), 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 re-impose this assumption once again.

With $G_L^{(j)}$, $j = 1,\dots,k$ as before, let

$$\hat\tau^*(L) = \min\Big(S^{(1)}_{G_L^{(1)}},\dots,S^{(k)}_{G_L^{(k)}}\Big). \tag{4.22}$$

Let us check first that

$$\lim_{L\to\infty} \bar F_{\rm on}\Big(\frac{L}{1-r}\Big)\Big(E\hat\tau^*(L) - E\tau^*(L)\Big) = 0. \tag{4.23}$$

We may assume that the random variables $\{X_i^{(j)},\ i\ge 1\}$, $\{X_i^{(*,j)},\ i\ge 1\}$, $\{Y_i^{(j)},\ i\ge 1\}$ and $G_L^{(j)}$, $j = 1,\dots,k$ are all defined on the same probability space, such that $X_i^{(j)} \ge X_i^{(*,j)}$ for all $i \ge 1$ and $j = 1,\dots,k$, and that the same $\{Y_i^{(j)},\ i\ge 1\}$ and $G_L^{(j)}$ are used to define $S^{(j)}_{G_L^{(j)}}$ and $\tau^{*,j}(L)$ for $j = 1,\dots,k$. We then have:

$$\hat\tau^*(L) - \tau^*(L) = \min\Big(S^{(1)}_{G_L^{(1)}},\dots,S^{(k)}_{G_L^{(k)}}\Big) - \min\Big(\tau^{*,1}(L),\dots,\tau^{*,k}(L)\Big) \le \bigvee_{1\le j\le k}\Big(S^{(j)}_{G_L^{(j)}} - \tau^{*,j}(L)\Big).$$

Since $S^{(j)}_{G_L^{(j)}} \ge \tau^{*,j}(L)$ for all $j = 1,\dots,k$, we have, therefore,

$$\begin{aligned}
E\hat\tau^*(L) - E\tau^*(L) \le& kE(S^{(1)}_{G_L^{(1)}} - \tau^{*,1}(L))\\
=& kEG_L^{(1)}(EX_1^{(1)} - EX_1^{(*,1)})\\
=& k\Big(\bar F_{\rm on}\Big(\frac{L}{1-r}\Big)^{-1} - 1\Big)o(1)\\
=& o\Big(\bar F_{\rm on}\Big(\frac{L}{1-r}\Big)^{-1}\Big)
\end{aligned}$$

as $L \to \infty$. This proves (4.23). Therefore, (4.21) will follow once we check that

$$\liminf_{L\to\infty} \bar{F}_{\rm on}\Big(\frac{L}{1-r}\Big) E\hat{\tau}^*(L) \geq \frac{1}{k}\mu. \tag{4.24}$$

Now,

$$E\hat{\tau}^*(L) = E\min\Big(S^{(1)}_{G^{(1)}_L}, \ldots, S^{(k)}_{G^{(k)}_L}\Big) = \int_0^\infty P\Big(S^{(1)}_{G^{(1)}_L} > x\Big)^k \, dx.$$

Choose an $N \geq 1$ (large) and $0 < \delta < 1$ (small) and observe that for any $x > 0$,

$$\begin{aligned}
P\Big(S^{(1)}_{G^{(1)}_L} > x\Big) =& P\Big(\sum_{i=1}^{G^{(1)}_L}(X^{(1)}_i + Y^{(1)}_i) > x\Big) \\
\geq& P\Big(\cap_{n>N}[G^{(1)}_L > \max(N, \frac{x}{(1-\delta)\mu}), \sum_{i=1}^n (X^{(1)}_i + Y^{(1)}_i) > n(1-\delta)\mu)]\Big) \\
=& P\Big(G^{(1)}_L > \max(N, \frac{x}{(1-\delta)\mu})\Big) P\Big(\cap_{n>N}[\sum_{i=1}^n (X^{(1)}_i + Y^{(1)}_i) > n(1-\delta)\mu)]\Big) \\
=:& \theta_N P\Big(G^{(1)}_L > \max(N, \frac{x}{(1-\delta)\mu})\Big).
\end{aligned} \tag{4.25}$$

Observe that by the strong law of large numbers, $\theta_N \to 1$ as $N \to \infty$. We have, therefore,

$$\begin{aligned}
E\hat{\tau}^*(L) \geq& \theta_N^k \int_{N(1-\delta)\mu}^\infty P\Big(G^{(1)}_L > \frac{x}{(1-\delta)\mu}\Big)^k \, dx \\
=& \theta_N^k \sum_{n=N}^\infty (1-\delta)\mu P\Big(G^{(1)}_L > n\Big)^k \\
=& \theta_N^k (1-\delta)\mu \frac{F_{\rm on}\Big(\frac{L}{1-r}\Big)^{k(N+1)}}{1 - F_{\rm on}\Big(\frac{L}{1-r}\Big)^k}.
\end{aligned}$$

We conclude that

$$\liminf_{L\to\infty} \bar{F}_{\rm on}\Big(\frac{L}{1-r}\Big) E\hat{\tau}^*(L) \geq \frac{1}{k}\mu\theta_N^k(1-\delta),$$

and (4.24) follows by letting $N \to \infty$ and $\delta \to 0$. Therefore, to prove the theorem we only need to prove (4.20).

By the assumptions of the theorem, there is a $p' > 1$ such that both $EX_1^{p'} < \infty$ and $EY_1^{p'} < \infty$. By Hölder's inequality,

$$E_{H,x_0}\Big(\tau^*(L(1-\epsilon))1(\tau(L) < \tau^*(L(1-\epsilon)))\Big) \leq \Big(E_{H,x_0}(\tau^*(L(1-\epsilon)))^{p'}\Big)^{1/p'} P_{H,x_0}(\tau(L) < \tau^*(L(1-\epsilon)))^{1/q'}, \tag{4.26}$$

where $\frac{1}{p'} + \frac{1}{q'} = 1$. We use the following inequality: for any σ-finite measure spaces $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$, a $p' > 1$ and a nonnegative measurable function $f : \Omega_1 \times \Omega_2 \to R$,

$$\Big(\int_{\Omega_1}\Big(\int_{\Omega_2} f(\omega_1,\omega_2)\, \mu_2(d\omega_2)\Big)^{p'} \mu_1(d\omega_1)\Big)^{1/p'} \leq \int_{\Omega_2}\Big(\int_{\Omega_1} f(\omega_1,\omega_2)^{p'} \, \mu_1(d\omega_1)\Big)^{1/p'} \mu_2(d\omega_2). \tag{4.27}$$

See for example Lemma 3.3.1 of Kwapień and Woyczyński (1992). Let $j_0 = \text{argmin}(D_1^{(0)}, \ldots, D_k^{(0)})$ (with ties broken in, say, the lexicographical manner), and recall that $D_{j_0}^{(0)} = 0$, P_{H,x_0}–almost surely. We have by (4.27)

$$\begin{aligned}
\Big(E_{H,x_0}(\tau^*(L(1-\epsilon)))^{p'}\Big)^{1/p'} \leq & \Big(E_{H,x_0}(\tau^*(L))^{p'}\Big)^{1/p'} \leq \Big(E_{H,x_0}(\hat{\tau}_L^*)^{p'}\Big)^{1/p'} \\
\leq & \Big(E_{H,x_0}(S^{(j_0)}_{G_L^{(j_0)}})^{p'}\Big)^{1/p'} \\
= & \left(E\Big(\sum_{i=1}^{\infty}(X_i^{(1)} + Y_i^{(1)})1(i \leq G_L^{(1)})\Big)^{p'}\right)^{1/p'} \\
\leq & \sum_{i=1}^{\infty}\left(E\Big((X_i^{(1)} + Y_i^{(1)})1(i \leq G_L^{(1)})\Big)^{p'}\right)^{1/p'} \\
= & \Big(E(X_1^{(1)} + Y_1^{(1)})^{p'}\Big)^{1/p'} \sum_{i=1}^{\infty} P(G_L^{(1)} \geq i)^{1/p'} \\
\leq & C\,\bar{F}_{\text{on}}\Big(\frac{L}{1-r}\Big)^{-1},
\end{aligned}$$

where C is a finite positive constant. It follows from (4.26) that we will prove (4.20) by showing that

$$\lim_{L\to\infty} P_{H,x_0}(\tau(L) < \tau^*(L(1-\epsilon))) = 0. \tag{4.28}$$

Let us prove first that for every

$$N > 2k\frac{\alpha}{\alpha - 1}, \tag{4.29}$$

we have

$$\lim_{L\to\infty} P_{H,x_0}(\tau(L) < \tau^*(L/N)) = 0. \tag{4.30}$$

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

$$dX_j(t) = Z_j(t)\ dt - \frac{1}{k}r(X_j)(t)\ dt,\ t \geq 0 \tag{4.31}$$

where $\{Z_j(t),\ t \geq 0\}$ is given by (4.7), and $X_j(0) = \frac{1}{k}x_0$, $j = 1, \ldots, k$. The k processes $\{X_j(t),\ t \geq 0\}$, $j = 1, \ldots, k$ are, conditionally on the initial delay $(D_1^{(0)} \ldots, D_k^{(0)})$, independent. Let $Y^{(k)}(t) = X_1(t) + \ldots + X_k(t)$, $t \geq 0$. The two processes, $\{X^{(k)}(t),\ t \geq 0\}$ and $\{Y^{(k)}(t),\ t \geq 0\}$ 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 rate of outflow of work from $Y^{(k)}(\cdot)$ does not exceed r. Therefore, for every ω,

$$X^{(k)}(t) \leq Y^{(k)}(t),\ t \geq 0. \tag{4.32}$$

Note that (4.32) is just an expression of the benefit of pooling the system resources. Define

$$\tau^{(Y)}(L) = \inf\{t \geq 0 :\ Y^{(k)}(t) \geq L\}. \tag{4.33}$$

Then (4.32) implies that $\tau^{(Y)}(L) \le \tau(L)$, so that

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

and, therefore,

$$P_{H,x_0}(\tau(L) < \tau^*(L/N)) \le P_{H,x_0}(\tau^{(Y)}(L) < \tau^*(L/N)). \tag{4.34}$$

Now, let for $j = 1, \dots, k$

$$\tau^{(j)}(L) = \inf\{t \ge 0:\ X_j(t) \ge L\}.$$

Then

$$\begin{aligned} P_{H,x_0}(\tau^{(Y)}(L) < \tau^*(L/N)) \le & P_{H,x_0}\Big(\tau^{(j)}(L/k) < \tau^*(L/N) \text{ for some } j = 1, \dots, k\Big) \\ & \le \sum_{i=1}^{j} P_{H,x_0}(\tau^{(j)}(L/k) < \tau^*(L/N)) \\ & \le \sum_{i=1}^{j} P_{H,x_0}(\tau^{(j)}(L/k) < \tau^{*,j}(L/N)). \end{aligned} \tag{4.35}$$

Therefore, (4.30) will follow from (4.34) and (4.35) once we prove that for every

$$M > 2\frac{\alpha}{\alpha - 1} \tag{4.36}$$

we have

$$\lim_{L\to\infty} P_{H,x_0}(\tau^{(j)}(L) < \tau^{*,j}(L/M)) = 0, \tag{4.37}$$

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

We will prove (4.37) for $j = 1$. Let $T_0 = \inf\{t > D_1^{(0)}:\ X_1(t) = 0\}$. Write

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

Since for all $L > 2x_0$

$$\begin{aligned} P_{H,x_0}(\tau^{(1)}(L) < \tau^{*,1}(L/M) \wedge T_0) \le & P_{H,x_0}(\tau^{(1)}(L) < T_0) \\ & \le P(\tau^{(1)}_{L/2} < T_0) \to 0 \end{aligned}$$

as $L \to \infty$ because of the rate condition (4.5) (or recall Corollary 2.2), (4.37) will follow if we prove that

$$\lim_{L\to\infty} P_{H,x_0}(T_0 \le \tau^{(1)}(L) < \tau^{*,1}(L/M)) = 0. \tag{4.38}$$

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. By the strong Markov property,

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

Therefore, (4.38) will follow once we prove (4.37) with $x_0 = 0$, and so (4.37) in its generality will follow if we prove it only for $x_0 = 0$. Assume, therefore, that $x_0 = 0$.

For $K_1 \geq 0$ and $K_2 > 0$ let $\{X_1^{(K_1,K_2)},\ t \geq 0\}$ denote the process given by (4.31) (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 $\{X_1^{(K_1,K_2)},\ t \geq 0\}$. Observe that the event

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

increases when K_1 and K_2 decrease. It is enough, therefore, to prove (4.37) 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 (4.37) 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 3 events, $A_i(L)$, $i = 1, 2, 3$, corresponding to the following 3 possibilities.

(*i*) $\tau^{*,1}(L/M) < \tau^{(1)}(L) \wedge T_0$. That is, the process $\{X_1(t),\ t \geq 0\}$ begins an *on* interval of length at least $\frac{L}{M(1-r)}$ before reaching either level L, or returning to 0.

(*ii*) $\tau^{(1)}(L) < \tau^{*,1}(L/M) \wedge T_0$. In other words, the process $\{X_1(t),\ t \geq 0\}$ reaches level L before starting an *on* interval of length at least $\frac{L}{M(1-r)}$ and before returning to 0.

(*iii*) $T_0 < \tau^{(1)}(L) \wedge \tau^{*,1}(L/M)$. In other words, the process $\{X_1(t),\ t \geq 0\}$ returns to 0 before reaching level L, and without having an *on* interval of the length of at least $\frac{L}{M(1-r)}$.

Clearly,

$$P(\tau^{(1)}(L) < \tau^{*,1}(L/M)) = P\Big(\{\tau^{(1)}(L) < \tau^{*,1}(L/M)\} \cap A_2(L)\Big) + P\Big(\{\tau^{(1)}(L) < \tau^{*,1}(L/M)\} \cap A_3(L)\Big).$$

However, by the strong Markov property,

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

Therefore,

$$P(\tau^{(1)}(L) < \tau^{*,1}(L/M)) = \frac{P\Big(\{\tau^{(1)}(L) < \tau^{*,1}(L/M)\} \cap A_2(L)\Big)}{1 - P(A_3(L))}. \tag{4.39}$$

Let $\xi_i = (1-r)X_i^{(1)} - rY_i^{(i)},\ i = 1, 2, \ldots$. Taking into account that the *off* times are bounded, we can repeat now the argument used to estimate $I_{11}(x)$ in the proof of Proposition 2.1′, to conclude that

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

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

$$1 - P(A_3(L)) \geq P(A_1(L)) \geq P\Big(X_1^{(1)} > \frac{L}{M(1-r)}\Big) = \bar{F}_{\text{on}}\Big(\frac{L}{M(1-r)}\Big).$$

We conclude by (4.39) and (4.36) that

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

This proves (4.37) and so (4.30) 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 (4.29)

$$\lim_{L\to\infty} \sup_{H} P_{H,0}(\tau(L) < \tau^*(L/N)) = 0. \tag{4.40}$$

The next step in the proof of (4.28) is to show that two "long" *on* periods are "unlikely to happen simultaneously". Formally, let

$$Q_L = \inf\{t \ge 0 : \text{at time } t \text{ there are two } on \text{ periods running, each of length at least } L\}. \tag{4.41}$$

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

$$\lim_{L\to\infty} P_{H,x_0}\left(Q_L \le \gamma_L^{-1}(\bar{F}_{\text{on}}(L))^{-1}\right) = 0. \tag{4.42}$$

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 $j_1, j_2 = 1, \dots, k$, $j_1 \ne j_2$, let

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

Then

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

and so for any $q > 0$,

$$P(Q_L \le q) \le \frac{k(k-1)}{2} P(Q_L^{(1,2)} \le q). \tag{4.43}$$

Let

$$\begin{aligned} Q_L^{(1)} = \inf\{&t \ge 0 : \text{at time } t, X_1(\cdot) \text{ begins an } on \text{ period of length at least } L \\ &\text{during which } X_2(\cdot) \text{ also begins an } on \text{ period of length at least } L\} \end{aligned} \tag{4.44}$$

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

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

which means that for any $q > 0$,

$$P(Q_L^{(1,2)} \le q) \le 2P(Q_L^{(1)} \le q). \tag{4.45}$$

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

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

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

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

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

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

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

$$p_L = P\Big(\inf\{S_n^{(2)} : n \ge 1,\ X_{n+1}^{(2)} \ge L\} \le W\Big). \tag{4.46}$$

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,

$$Q_L^{(1)} \stackrel{\rm st}{\ge} \sum_{n=1}^{H_L} Z_n. \tag{4.47}$$

We claim that

$$p_L \to 0 \text{ as } L \to \infty. \tag{4.48}$$

Indeed,

$$p_L = \frac{1}{\bar{F}_{\rm on}(L)} \int_L^\infty P(Z_1 \le t)\ F_{\rm on}(dt) = \frac{1}{\bar{F}_{\rm on}(L)} \int_L^\infty P(\sum_{i=1}^{G_L} \hat{X}_i \le t)\ F_{\rm on}(dt). \tag{4.49}$$

Observe that for every $t \ge L$,

$$\begin{aligned} P(\sum_{i=1}^{G_L} \hat{X}_i \le t) \le & P(G_L \le \frac{1}{(\bar{F}_{\rm on}(L))^{1/2}}) + P\Big(\inf_{n > (\bar{F}_{\rm on}(L))^{-1/2}} \sum_{i=1}^{n} \hat{X}_i - n\frac{\mu_{\rm on}}{2} \le 0\Big) \\ & + P(\frac{\mu_{\rm on}}{2} G_L \le t). \end{aligned} \tag{4.50}$$

Notice that the first term in the right hand side of (4.50) goes to 0 as $L \to \infty$, and that by the strong law of large numbers so does the second term in the the right hand side of (4.50). Since neither of these two terms depends on t, (4.48) will follow once we prove that

$$\lim_{L \to \infty} \frac{1}{\bar{F}_{\rm on}(L)} \int_L^\infty P(\frac{\mu_{\rm on}}{2} G_L \le t)\ F_{\rm on}(dt) = 0. \tag{4.51}$$

However, for all L big enough and all $t \geq L$ we have $[2t/\mu_{\rm on}] \geq 1$, and so

$$P(\frac{\mu_{\rm on}}{2} G_L \leq t) = 1 - \Big(F_{\rm on}(L)\Big)^{[2t/\mu_{\rm on}]} \leq \Big[\frac{2t}{\mu_{\rm on}}\Big] \bar{F}_{\rm on}(L).$$

Therefore,

$$\limsup_{L\to\infty} \frac{1}{\bar{F}_{\rm on}(L)} \int_L^\infty P(\frac{\mu_{\rm on}}{2} G_L \leq t)\, dt \leq \frac{2}{\mu_{\rm on}} \limsup_{L\to\infty} \int_L^\infty t\ F_{\rm on}(dt) = 0$$

since $\mu_{\rm on} < \infty$. This proves (4.51), and, therefore, we have (4.48).

Define now $\gamma_L = p_L^{1/2}$. Observe that by (4.43) (4.45) and (4.47) we have

$$P\Big(Q_L \leq \gamma_L^{-1} (\bar{F}_{\rm on}(L))^{-1}\Big) \leq k(k-1) P\Big(\sum_{n=1}^{H_L} Z_n \leq p_L^{-1/2} (\bar{F}_{\rm on}(L))^{-1}\Big). \tag{4.52}$$

However,

$$\sum_{n=1}^{H_L} Z_n \stackrel{d}{=} \sum_{i=1}^{J_L} \hat{X}_i,$$

where J_L is a geometric random variable with parameter $p_L \bar{F}_{\rm on}(L)$, independent of $\{\hat{X}_i,\ i = 1, 2, \dots\}$. Since, as $L \to \infty$,

$$(p_L \bar{F}_{\rm on}(L)) J_L \Rightarrow E(1)$$

where $E(1)$ is a standard exponential random variable and we immediately obtain (4.42) using (4.52) and the law of large numbers.

Let us go back now to the proof of (4.28). Observe, first of all, that for all L big enough,

$$P_{H,x_0}\Big(\tau(L) < \tau^*(L(1-\epsilon))\Big) \leq P_{H,0}\Big(\tau(L(1-\epsilon/2)) < \tau^*(L(1-\epsilon))\Big).$$

Therefore, it is enough to prove (4.28) 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 (4.29) and big enough to make the right hand side of (4.53) below positive, and observe that it is enough to prove (4.28) for $\epsilon = 1/N$. Fix, further, a ρ satisfying

$$(k-1)\rho \leq \frac{\epsilon}{2} - \epsilon^2. \tag{4.53}$$

We have

(4.54)

$$\begin{aligned} P_H(\tau(L) < \tau^*(L(1-\epsilon))) \leq & P_H\Big(\tau(L) < \tau^*(L(1-\epsilon)), \tau(L\epsilon^2) \geq \tau^*(L\epsilon^3), \tau(L) < Q_{\rho L}\Big) \\ & + P_H(\tau(L\epsilon^2) < \tau^*(L\epsilon^3)) + P_H(\tau(L) \geq Q_{\rho L}). \end{aligned}$$

Observe that by (2.30)

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

Furthermore, by (4.42) and (4.18)

$$\begin{aligned} P_H(\tau(L) \geq Q_{\rho L}) \leq & P_H\Big(Q_{\rho L} \leq (\gamma_{\rho L})^{-1} (\bar{F}_{\rm on}(\rho L))^{-1}\Big) + P_H\Big(\tau(L) \geq (\gamma_{\rho L})^{-1} (\bar{F}_{\rm on}(\rho L))^{-1}\Big) \\ \leq & P_H\Big(Q_{\rho L} \leq (\gamma_{\rho L})^{-1} (\bar{F}_{\rm on}(\rho L))^{-1}\Big) + \gamma_{\rho L} \bar{F}_{\rm on}(\rho L) E_H \tau(L) \\ \to & 0 \end{aligned}$$

as $L \to \infty$. Therefore, (4.28) will follow once we prove that

$$\lim_{L\to\infty} P_H\Big(\tau(L) < \tau^*(L(1-\epsilon)),\ \tau(L\epsilon^2) \geq \tau^*(L\epsilon^3),\ \tau(L) < Q_{\rho L}\Big) = 0. \tag{4.55}$$

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

$$\lim_{L\to\infty} \sup_H P_H\Big(\tau(L) < \tau^*(L(1-\epsilon)),\ \tau(L\epsilon^2) \geq \tau^*(L\epsilon^3),\ \tau(L) < Q_{\rho L}\Big) = 0. \tag{4.56}$$

Let

$$B(L) = \Big\{\tau(L) < \tau^*(L(1-\epsilon)),\ \tau(L\epsilon^2) \geq \tau^*(L\epsilon^3),\ \tau(L) < Q_{\rho L}\Big\}.$$

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

(i) After starting the first "wet period" the process $X^{(k)}(\cdot)$ reaches level L before returning to 0.

(ii) 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 $\tau(L\epsilon^2)$ 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 $\tau(L\epsilon^2)$ of length least ρL, has length not exceeding $\frac{L(1-\epsilon)}{(1-r)}$. Let us now modify the state of the system at time $\tau(L\epsilon^2)$ in the following way. Bring all the work remaining in the presently running *on* periods in one "lump" at time $\tau(L\epsilon^2)$, and attach the remainders of these *on* periods 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

$$L\epsilon^2 + L(1-\epsilon) + (k-1)\rho L \leq L(1-\epsilon/2)$$

by (4.53). We increase, if necessary, the state of the system to exactly $L(1-\epsilon/2)$ (only increasing in the process the probability of the event $B_{12}(L)$). We conclude that for some H_0,

$$P_H(B_{12}(L)) \leq P_{H_0, L(1-\epsilon/2)}\Big(\sup_{t\geq 0} V^{(k)}(t) \geq L\Big), \tag{4.57}$$

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

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

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

$$V^{(k)}(t) = V_1(t) + \ldots + V_k(t),\ t \geq 0,$$

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

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

with $V_j(0) = L(1-\epsilon/2)/k$, and with initial delay governed by the jth marginal law, $H^{(j)}$, of H_0. We conclude immediately by (4.57) that

$$P_H(B_{12}(L)) \leq k \sup_{H^{(1)}} P_{H^{(1)}, L(1-\epsilon/2)/k}(\sup_{t\geq 0} V_1(t) \geq \frac{L}{k}) = kP(\sup_{t\geq 0} V_1(t) \geq \frac{L\epsilon}{2k}),$$

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

$$\lim_{L\to\infty} \sup_H P_H(B_{12}(L)) = 0. \tag{4.58}$$

In exactly the same way one can show that

$$\lim_{L\to\infty} \sup_H P_H(B_{11}(L)) = 0. \tag{4.59}$$

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 $L\epsilon^2$ before the beginning of the first *on* period of length at least $L\epsilon^3$, or not. Accordingly, by the strong Markov property,

$$P_H(B_2(L)) \le P_H(\tau^*(L\epsilon^3) < \tau^*(L(1-\epsilon)))\Big(\sup_G P_G(\tau(L\epsilon^2) \le \tau^*(L\epsilon^3)) + \sup_G P_G(B(L))\Big).$$

Taking supremum over H, we obtain

$$\begin{aligned}\sup_H P_H(B(L)) \le &\sup_H P_H(B_{11}(L)) + \sup_H P_H(B_{12}(L)) \\ &+ P(\tau^*(L\epsilon^3) < \tau^*(L(1-\epsilon)))\Big(\sup_H P_H(\tau(L\epsilon^2) \le \tau^*(L\epsilon^3)) + \sup_H P_H(B(L))\Big),\end{aligned}$$

which is the same as

$$\sup_H P_H(B(L)) \le \frac{\sup_H P_H(B_{11}(L)) + \sup_H P_H(B_{12}(L)) + \sup_H P_H(\tau(L\epsilon^2) \le \tau^*(L\epsilon^3))}{P(\tau^*(L\epsilon^3) = \tau^*(L(1-\epsilon)))}. \tag{4.60}$$

Now (4.56) follows from (4.60), (4.58), (4.59), (4.40) and the fact that

$$P(\tau^*(L\epsilon^3) = \tau^*(L(1-\epsilon))) = \frac{\bar F_{\rm on}(L(1-\epsilon))}{\bar F_{\rm on}(L\epsilon^3)} \to \Big(\frac{\epsilon^3}{1-\epsilon}\Big)^\alpha > 0$$

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

5. Appendix.

We will construct explicitly a stationary version of $\{X^{(k)}(t),\ t \ge 0\}$. In fact, we will construct a stationary version of $\{X^{(k)}(t),\ -\infty < t < \infty\}$ on the whole real line. We proceed as follows. Let $\{Z_j(t),\ -\infty < t < \infty\},\ j = 1, \dots, k$ be stationary. Then $\{Z^{(k)}(t),\ -\infty < t < \infty\}$ defined by (4.3) is itself stationary. We will construct $\{X^{(k)}(t),\ -\infty < t < \infty\}$ by exhibiting a measurable function $\varphi:\ D(\mathbb{R}, \{0, 1, \dots, k\}) \to \mathbb{R}_+$ such that the process

$$X^{(k)}(t) = \varphi(Z^{(k)}(\cdot + t)),\ -\infty < t < \infty \tag{5.1}$$

satisfies (4.4). Clearly, the process given by (5.1) is stationary.

Let $\mathbf{x} = (x(u),\ -\infty < u < \infty) \in D(\mathbb{R}, \{0, 1, \dots, k\})$, and let

$$\dots < T_{-2} < T_{-1} < 0 \le T_1 < T_2 < \dots$$

be the epoch times for $\{x(u),\ -\infty < u < \infty\}$. That is, at each time T_i, $\{x(u),\ -\infty < u < \infty\}$ changes its value, and let

$$\mathcal{Z}_{0,1} = \{i:\ x(T_i - 0) = 0,\ x(T_i + 0) = 1\}.$$

For $\mathbf{x} = (Z^{(k)}(t),\ -\infty < t < \infty)$, the points T_i with $i \in \mathcal{Z}_{0,1}$ are the times when one of the *on/off* processes begins an *on* period, following a period when all k processes were off. For every $i \in \mathcal{Z}_{0,1}$ let L_i be the length of the "wet period" commencing at T_i. Formally, let $X^{(k)}(T_i) = 0$, and define $\{X^{(k)}(t),\ u \geq T_i\}$ by the following analogue of (4.4)

$$dX^{(k)}(u) = x(u)\ du - r(X^{(k)})(u)\ du. \tag{5.2}$$

Then

$$L_i = \inf\{u > 0 : X^{(k)}(T_i + u) = 0\}.$$

If we denote

$$G_1 = \Big\{\mathbf{x} :\ \text{ for every } i \in \mathcal{Z}_{0,1}(\mathbf{x}),\ L_i(\mathbf{x}) < \infty\Big\},$$

then the rate condition (4.5) implies that the set $G_1 \in D(\mathbb{R}, \{0, 1, \ldots, k\})$ has full measure under the probability measure μ_0 induced by $\{Z^{(k)}(t),\ -\infty < t < \infty\}$ on $D(\mathbb{R}, \{0, 1, \ldots, k\})$. Now, for a $-\infty < t < \infty$ let

$$\mathcal{A}_{0,1}(t) = \{T_i \leq t :\ i \in \mathcal{Z}_{0,1}\},$$

and define

$$T(t) = \inf\{T_i \in \mathcal{A}_{0,1}(t) : T_i + L_i \geq t\}. \tag{5.3}$$

Denote

$$G_2 = \Big\{\mathbf{x} :\ \text{for every } -\infty < t < \infty,\ T_i + L_i \geq t \text{ for only finitely many } i \in \mathcal{A}_{0,1}(t)\Big\}.$$

We claim that the set $G_2 \in D(\mathbb{R}, \{0, 1, \ldots, k\})$ has full measure under the probability measure μ_0. That is, we claim that

$$P\Big(\Big\{\omega :\ \text{for some } -\infty < t < \infty,\ T_i + L_i \geq t, \text{ for infinitely many } i \in \mathcal{A}_{0,1}(t)(Z^{(k)}(\cdot)(\omega))\Big\}\Big) = 0. \tag{5.4}$$

Now, concentrating on the rational times only, we see that (5.4) will follow once we prove that for any $-\infty < t < \infty$ we have

$$P\Big(\Big\{\omega : T_i + L_i > t \ \text{ for infinitely many } i \in \mathcal{A}_{0,1}(t)(Z^{(k)}(\cdot)(\omega))\ \Big\}\Big) = 0. \tag{5.5}$$

We check (5.5) for $t = 0$. We have

$$P\Big(\Big\{\omega : T_i + L_i > 0 \text{ for infinitely many } i \in \mathcal{A}_{0,1}(0)(Z^{(k)}(\cdot)(\omega))\ \Big\}\Big)$$
$$\leq P\Big(\limsup_{n \to -\infty}\{\omega :\ \sum_{j=1}^{k} \int_n^0 Z_j(u)\ du \geq r|n| - r\}\Big)$$

and the last probability goes to 0 as $n \to -\infty$ because of the rate condition (4.5) and the obvious fact that

$$\frac{1}{|n|} \sum_{j=1}^{k} \int_n^0 Z_j(u)\ du \to k\frac{\mu_{\text{on}}}{\mu}$$

as $n \to -\infty$. This proves (5.5) and so (5.4) holds.

Finally, let

$$G_3 = \Big\{\mathbf{x} :\ \text{ for every } -\infty < t < \infty, \text{ there is a } u \leq t \text{ such that } x(u) = k\Big\}.$$

Note that $F_{\text{on}} * F_{\text{off}}$ being non-arithmetic implies that the set $G_3 \in D(\mathbb{R}, \{0, 1, \ldots, k\})$ has full measure under the probability measure μ_0. Finally, we let $G = G_1 \cap G_2 \cap G_3$, and we note that G has also full measure under the probability measure μ_0.

We will define now the function $\varphi(\mathbf{x})$, $\mathbf{x} \in D(\mathbb{R}, \{0, 1, \ldots, k\})$. If $\mathbf{x} \in G^c$, let $\varphi(\mathbf{x}) = 0$ (an arbitrary number). Suppose now that $\mathbf{x} \in G$. Since $\mathbf{x} \in G_3$, we may meaningfully define $s(\mathbf{x}) = \sup\{u \leq 0 : x(u) = k\}$. Since $\mathbf{x} \in G_2$, $T(\mathbf{x}, s(\mathbf{x}))$ is well defined by (5.3). Now define $\varphi(\mathbf{x})$ as $X^{(k)}(0)$ given by solving (5.2) on $(T(\mathbf{x}, s(\mathbf{x})), \infty)$, with $X^{(k)}(T(\mathbf{x}, s(\mathbf{x}))) = 0$. Intuitively, $s(Z^{(k)}(\cdot)))$ is the last time before time 0 that all k *on/off* processes were on, which is identifiable as a time point when the system is in a "wet period", and $T(Z^{(k)}(\cdot), s(Z^{(k)}(\cdot)))$ is the beginning of that "wet period", at which time the system must be empty. This constructs the required function $\varphi : D(\mathbb{R}, \{0, 1, \ldots, k\}) \to \mathbb{R}_+$, and thus a stationary version of a process satisfying (3.4).

References

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

Anick, D., Mitra, D. and Sondhi, M., *Stochastic theory of a data-handling system with multiple sources*, Bell System Tech. J. **61** (1982), 1871–1894.

Asmussen,S., *Applied Probability and Queues*, J. Wiley and Sons, New York, 1987.

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

Avram, Florin and Taqqu, Murad S., *Weak convergence of moving averages with infinite variance*, Dependence in Probability and Statistics, Progr. Probab. Statist., 11, Birkhauser Boston, Boston, MA, 1986, pp. 399–415.

Avram, Florin and Taqqu, Murad S., *Probability bounds for M-Skorohod oscillations*, Stochastic Processes and their Applications **33** (1989), 63–72.

Avram, Florin and Taqqu, Murad S., *Weak convergence of sums of moving averages in the α-stable domain of attraction*, Ann. Probability **20** (1992), 483–503.

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

Billingsley, P., *Convergence of Probability Measures*, Wiley, New York, 1968.

Bingham, N., Goldie, C. and Teugels, J., *Regular Variation*, Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, UK, 1987.

Brichet, F., Roberts, J., Simonian, A. and Veitch, D., *Heavy traffic analysis of a storage model with long range dependent on/off sources*, Preprint (1996).

Choudhury, G. and Whitt, W., *Long-tail buffer-content distributions in broadband networks*, Preprint, AT&T Bell Laboratories (1995).

Crovella, M and Bestavros, A., *Explaining world wide web traffic self–similarity*, Preprint available as TR-95-015 from {crovella,best}@cs.bu.edu (1995).

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

Feller, W., *An Introduction to Probability Theory and its Applications*, Vol. II, second edition, Wiley, New York, 1971.

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

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

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

Kwapien, S. and Woyczynski, N.A., *Random Series and Stochastic Integrals: Single and Multiple*, Birkhauser, Boston, 1992.

Leland, W., Taqqu, M., Willinger, W., Wilson, D., *On the self–similar nature of ethernet traffic*, Proceedings of the ACM/SIGCOMM '93, San Francisco, ACM/SIGCOMM Computer Communications Review **23** (1993), 183–193.

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

Pacheco, A. and Prabhu, N.U., *A Markovian storage model*, Ann. Applied Probability **6** (1996), 76–91.

Prabhu, N.U. and Pacheco, A., *A storage model for data communication systems*, Queueing Systems **19** (1995), 1–40.

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

Resnick, Sidney, *Point processes, regular variation and weak convergence*, Advances Appl. Probability **18** (1986), 66-138.

Resnick, Sidney, *Extreme Values, Regular Variation, and Point Processes*, Springer-Verlag, New York, 1987.

Resnick, S., *Adventures in Stochastic Processes*, Birkhauser, Boston, 1992.

Willinger, W., Taqqu, M., Leland, W. and Wilson, D., *Self–similarity in high–speed packet traffic: analysis and modeling of ethernet traffic measurements*, Statistical Science **10** (1995), 67–85.

Willinger, W., Taqqu, M., Sherman, R. and Wilson, D., *Self–similarity through high–variability: Statistical analysis of ethernet LAN traffic at the source level*, Preprint (1995).

David Heath, Cornell University, School of Operations Research and Industrial Engineering, ETC Building, Ithaca, NY 14853 USA
E-mail address: davidh@orie.cornell.edu

Sidney I. Resnick, Cornell University, School of Operations Research and Industrial Engineering, ETC Building, Ithaca, NY 14853 USA
E-mail address: sid@orie.cornell.edu

Gennady Samorodnitsky, Cornell University, School of Operations Research and Industrial Engineering, ETC Building, Ithaca, NY 14853 USA
E-mail address: gennady@orie.cornell.edu