BUFFER CONTENT OF A LEAKY BUCKET SYSTEM WITH LONG–RANGE DEPENDENT INPUT TRAFFIC

BÁRBARA GONZÁLEZ–ARÉVALO AND GENNADY SAMORODNITSKY

ABSTRACT. Leaky bucket is a flow control mechanism that is designed to reduce the effect of the inevitable variability in the input stream into a node of a communication network. In this paper we study what happens when an input stream with heavy tailed work sessions arrives to a server protected by such a leaky bucket. Heavy tailed sessions produce long range dependence in the input stream. Previous studies of the systems without flow control suggested that such long range dependence can have dramatic effect on the system performance. By concentrating on the expected time till overflow of a large finite buffer we show that leaky bucket flow control does make the system overflow less often, but long range dependence still makes its presence felt.

1. INTRODUCTION

The study of traffic on data networks has changed substantially since the appearance of modern communication systems, which are essentially different from the traditional voice traffic networks. The main difference that appears in modern networks is the dependence structure of the data. While traditional models are based on assumptions of short range dependence, recent measurements (see Leland et al. (1994), Paxson and Floyd (1994), Cunha et al. (1995), Crovella and Bestavros (1996)) show the presence of long–range dependence and self–similarity in the data of network traffic. Presently it is believed that these phenomena are caused by the presence of heavy tails in the distribution of the service times, which cause the long–range dependence. We consider a fluid version of a leaky bucket flow control protocol, with an input process in which the distribution of the session lengths is heavy tailed, causing it to be long–range dependent. We will consider two types of input processes: an On–Off process and a Poisson process. Recently there has been a lot of work concerning fluid models fed by On–Off or Poisson processes (see, for example, Heath et al. (1997, 1999), Jelenković and Lazar (1999), a survey in Boxma and Dumas (1998) and a recent study in Zwart et al. (2000)). The main concerns of these studies have been motivated by design and performance issues, but most of these studies ignore the fact that actual networks usually have some kind of policing mechanism (like TCP or the leaky bucket). In this paper we concentrate on certain design and performance issues related to the presence of a specific policing mechanism: the leaky bucket. Queuing systems with such control mechanism have been studied before, in particular in a series of papers of A. Berger and W. Whitt

1991 *Mathematics Subject Classification.* Primary 90B15; secondary 90B18, 60K25.

Key words and phrases. fluid queue, flow control, leaky bucket, finite buffer, heavy tails, long range dependence, time until overflow .

Research partially supported by NSF grant DMS-0071073 at Cornell University. Samorodnitsky's research was also partially supported by NSF grant DMI-9713549 at Cornell University.

(Berger and Whitt (1992c,b,a, 1994)). However, to the best of our knowledge only the paper Vamvakos and Anantharam (1998) looked at how the leaky bucket input control performs in the presence if a long range dependent input. Their conclusion was that the leaky bucket input control does not eliminate long range dependence. The general message from the results in the present paper is similar: long range dependence in the input stream still affects the system performance even when the leaky bucket input control is present. However, while Vamvakos and Anantharam (1998) concentrated only on the rate of decay of correlations, we look directly at system performance, specifically at the expected time until overflow of a large buffer. We show that the buffer still overflows much more often then in the "classical" case, without heavy tailed sessions, hence long range dependent input. In spite of that the leaky bucket input control will reduce the frequency at which the buffer overflows, in comparison with a system with the same input stream but without input control. We should also mention that, unlike the previous authors, who looked at discrete time systems, we investigate a fluid-type, continuous time system.

This paper is organized as follows. In section 2 we describe the system in detail and all the assumptions we are making about the parameters and the processes involved, and in section 3 we calculate the asymptotic expected time until a buffer of finite capacity overflows.

2. Description of the system

Consider a model of a network server with a leaky bucket policing mechanism defined as follows. Work arrives to the system according to some input process. We are going to consider two types of input processes: an On–Off process and a Poisson-type, or $M/G/\infty$ type process. For the On–Off process each session lasts a random length of time. The distribution of an On session's length is F_{on} and the distribution of an Off session's length is F_{off}. Both distributions have finite mean: μ_{on} and μ_{off} respectively. The lengths of different sessions are independent of each other. In the second case we are going to consider sessions arriving according to a Poisson process with rate $\lambda > 0$. Each session lasts a random length of time with distribution F that has a finite mean μ. The lengths of different sessions are independent of each other and of the Poisson arrival process.

In both cases a session generates work at unit rate. This work arrives at the infinite buffer of the leaky bucket. The departure of work from this buffer is controlled by *tokens* that arrive at a buffer of fixed size C at rate γ. Arriving work can be transmitted instantaneously to the server by consuming tokens. If the token buffer is empty, the work has to wait for the generation of new tokens. Stored work is transmitted immediately upon the generation of new tokens. The work that cannot be processed immediately by the server is collected in a buffer. The server is capable of processing $r > 0$ units of work per unit of time.

This system can be described by the following system of equations:

$$
\begin{aligned}
dX(t) &= (E(t) - r\mathbb{1}_{(X(t)>0)})dt \\
dY(t) &= (N(t) - E(t))dt \\
dZ(t) &= (\gamma\mathbb{1}_{(Z(t)<C)} - E(t))dt \\
E(t) &= N(t)\mathbb{1}_{(Z(t)>0)} + \left(\gamma\mathbb{1}_{(Y(t)>0)} + \min(N(t),\gamma)\mathbb{1}_{(Y(t)=0)}\right)\mathbb{1}_{(Z(t)=0)},
\end{aligned}
\tag{2.1}
$$

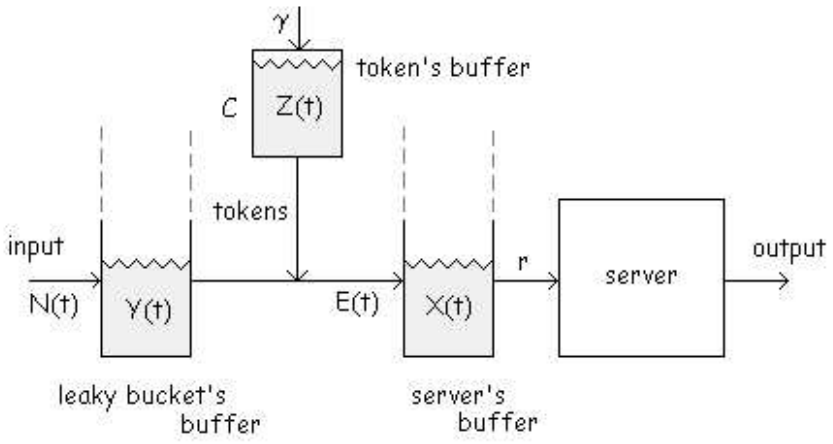

FIGURE 1. Fluid version of a leaky bucket flow control protocol.

where $Y(t)$ is the leaky bucket's buffer content, $Z(t)$ is the token's buffer content, $X(t)$ is the server's buffer content and $N(t)$ is the input process at time $t \geq 0$. Finally, $E(t)$ is the instantaneous rate at which work moves from the leaky bucket's buffer to the server's buffer. In the On–Off case $N(t)$ is 1 during an On session and 0 otherwise, and in the Poisson case $N(t)$ is the number of sessions running at time $t \geq 0$. Note that in the Poisson case $N(t)$ can be viewed as the number of customers in the system at time t in a $M/G/\infty$ queue, in which the sessions are customers and their lengths are their job requirements.

Assume that, in the On–Off case, the session length distribution for the On periods has a *regularly varying tail*. That is,

$$1 - F_{\text{on}}(x) = x^{-\alpha_{\text{on}}} L_{\text{on}}(x), \text{ as } x \to \infty,$$

where L_{on} is a slowly varying function, and $\alpha_{\text{on}} > 1$.

A function is said to be *slowly varying* if

$$\lim_{x \to \infty} \frac{L(ax)}{L(x)} = 1, \text{ for all } a > 0.$$

This assumption is a common way to model heavy tails of session lengths. The assumption $\alpha_{\text{on}} > 1$ assures finite mean session lengths (but sometimes infinite variance, when $\alpha_{\text{on}} < 2$) and hence makes it possible for the system to be stable if the service rate r is high enough. In the Poisson case assume that

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

where L is a slowly varying function, and $\alpha > 1$.

In the Poisson-type case we assume from now on that

$$0 < \lambda\mu < r < \gamma < 1 \tag{2.2}$$

and in the On–Off case, if we let $\theta := \dfrac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}}$, then

$$0 < \theta < r < \gamma < 1. \tag{2.3}$$

The first part of these assumptions is a way of making sure that there is a stationary version of the process, since for this we need the mean amount of work arriving into the system ($\lambda\mu$ or θ) to be less than the rate at which the server works (r). The assumption $\gamma > r$ assures that the server is never idle when there is work to be done, which seems like a reasonable assumption in a system of this kind. The assumption $\gamma < 1$ assures that when one session is running the content of the leaky bucket's buffer immediately begins to increase, so one session is enough to change the direction of the drift. In the Poisson-type case one can conjecture, based on previous studies of these systems without any flow control mechanism (Heath et al. (1999), Zwart et al. (2000)) that similar results may be expected in the case $\gamma < 1 + \lambda\mu$, since when one long session is running the other sessions bring in work at rate $\lambda\mu$.

It can be shown that the five–dimensional process describing the state of the system, $\{(X(t), Y(t), Z(t), N(t), E(t)), t \geq 0\}$, turns out to be a nice regenerative process based on a non–terminating renewal process with finite mean interarrival times. In particular, the process has a stationary distribution. We do not pursue this point here since the initial distribution of the system does not affect the expected time until overflow.

3. Expected time until buffer overflow

The following is our main theorem. It shows that under our assumptions the rate at which expected time until buffer overflow grows, roughly, $1/\big(1 - F_{\rm on}(H)\big)$ in the case of the On–Off input and $1/\big(1 - F(H)\big)$ in the case of Poisson input, where H is the size of the buffer.

Theorem 3.1. *For the On–Off input process, if for some $p > 1$ the p-th moment exists for the Off session length distribution, then*

$$\lim_{H\to\infty} H^{-\alpha_{on}} L_{on}(H) E\tau(H) = (\mu_{on} + \mu_{off}) \left(\frac{\gamma - \theta}{(\gamma - r)(1 - \theta)}\right)^{\alpha_{on}}. \tag{3.1}$$

For the Poisson input process,

$$\lim_{H\to\infty} H^{-\alpha} L(H) E\tau(H) = \frac{1}{\lambda} \left(\frac{\gamma - \lambda\mu}{\gamma - r}\right)^{\alpha}. \tag{3.2}$$

Remark 3.1. The results of Theorem 3.1 should be compared to the corresponding performance results without the leaky bucket input control. Then

$$\lim_{H\to\infty} H^{-\alpha_{\rm on}} L_{\rm on}(H) E\tau(H) = (\mu_{\rm on} + \mu_{\rm off}) \left(\frac{1}{1 - r}\right)^{\alpha_{\rm on}}$$

for the On–Off input process (see Theorem 2.3 of Heath et al. (1997)) and

$$\lim_{H\to\infty} H^{-\alpha} L(H) E\tau(H) = \frac{1}{\lambda} \left(\frac{1}{1 - r + \lambda\mu}\right)^{\alpha}$$

for the Poisson input process; see Proposition 4 and the subsequent comment in Resnick and Samorodnitsky (1999). One can immediately see that, while the leaky bucket input control does not change the order of magnitude at which the expected time until buffer overflow grows, it does make this expected time longer.

Furthermore, since the expression in the right hand sides of (3.1) and (3.2) are decreasing functions of γ, the buffer will overflow less often if γ is chosen as close to r as possible, which is entirely consistent with the logic of flow control.

Before proving Theorem 3.1 we establish some lemmas that are going to be used to prove the main result.

Lemma 3.1. *If $A(t) = \int_0^t N(s)ds$ is the amount of work received by the system in a time interval of length t with the On–Off or Poisson input process, then*

$$\frac{A(l)}{l} \to \begin{cases} \theta & \text{for the On–Off input} \\ \lambda\mu & \text{for the Poisson input} \end{cases} \text{ a.s. as } l \to \infty.$$

Proof.

We start with the Poisson case. We may assume, without loss of generality, that $N(0) = 0$. Consider the following regeneration times of the process $\{N(t), t \geq 0\}$:

$$N_0 = 0,$$

$$N_i = \inf\left\{ t > N_{i-1} : N(t) = 0, \sup_{N_{i-1} \leq s \leq t} N(s) > 0 \right\}, \text{ for } i \geq 1.$$

These times are the ends of busy periods in an $M/G/\infty$ queue with a finite mean service time distribution and, hence, $EN_1 < \infty$.

Now define the following random variables

$$A_i = A(N_i) - A(N_{i-1}) = \int_{N_{i-1}}^{N_i} N(s)ds, \text{ for } i \geq 1.$$

These random variables are iid since $\{N_i, i \geq 0\}$ are renewal times. Let $\{a_i\}$ be the arrival times of the Poisson process (that is, a_i is an Erlang random variable with i degrees of freedom, $i \geq 1$), and let l_i be the length of the session arriving at time a_i. Now consider the following filtration, where

$$\mathcal{F}_n = \sigma\Big(l_1, \ldots, l_n, a_1, \ldots, a_n, a_{n+1}\Big), \ n = 0, 1, \ldots .$$

Then

$$M_1 = \inf\{i \geq 1 : \ a_{i+1} > a_j + l_j, \ j = 1, \ldots, i\}$$

is a stopping time with respect to that filtration, and by Wald's lemma $EM_1 = \lambda E N_1 < \infty$. So, using Wald's lemma once again, we obtain

$$EA_1 = E\left(\int_0^{N_1} N(s)ds\right) = E\left(\sum_{i=1}^{M_1} l_i\right) = EM_1 El_1 = \lambda\mu E N_1.$$

By the law of large numbers we have that

$$\frac{1}{nEN_1} \sum_{i=1}^{n} A_i \to \lambda\mu, \text{ a.s. as } n \to \infty. \tag{3.3}$$

Let N_k be the biggest of the renewal times less than or equal to l. Then we have that, as $l \to \infty$,

$$\lambda\mu = \lim_{l\to\infty} \frac{1}{kEN_1} \sum_{i=1}^{k} A_i = \lim_{l\to\infty} \frac{l}{kEN_1} \frac{1}{l} \int_0^{N_k} N(s)ds \leq$$

$$\leq \liminf_{l\to\infty} \frac{1}{l} \int_0^{l} N(s)ds = \liminf_{l\to\infty} \frac{A(l)}{l} \leq \limsup_{l\to\infty} \frac{A(l)}{l} = \limsup_{l\to\infty} \frac{1}{l} \int_0^{l} N(s)ds \leq \tag{3.4}$$

$$\leq \lim_{l\to\infty} \frac{l}{(k+1)EN_1}\frac{1}{l}\int_0^{N_{k+1}} N(s)ds = \lim_{l\to\infty} \frac{1}{(k+1)EN_1}\sum_{i=1}^{k+1} A_i = \lambda\mu,$$

since $\dfrac{l}{kEN_1} \to 1$ as $l \to \infty$, and $N(s) \geq 0$ for all $s \geq 0$. So we finally get that

$$\frac{A(l)}{l} \to \lambda\mu, \text{ a.s. as } l \to \infty.$$

Similarly, in the On–Off case we define N_i to be the time when the i-th On session starts, $i = 0, 1, \ldots$, and $N_0 = 0$. Now define the following random variables

$$A_i = A(N_i) - A(N_{i-1}) = \int_{N_{i-1}}^{N_i} N(s)ds, \text{ for } i \geq 1.$$

These random variables are iid since the lengths of the On and Off sessions are all independent and the On sessions have all the same session length distribution as do the Off sessions. Moreover,

$$EA_1 = E\left(\int_0^{N_1} N(s)ds\right) = \mu_{\text{on}} = \theta(\mu_{\text{on}} + \mu_{\text{off}}).$$

So, by the law of large numbers we have that

$$\frac{1}{n(\mu_{\text{on}} + \mu_{\text{off}})}\sum_{i=1}^{n} A_i \to \theta, \text{ a.s. as } n \to \infty. \tag{3.5}$$

Now the remaining part of the lemma follows from (3.5) in the same way as its first part followed from (3.6).

Q.E.D.

Note that this lemma implies that for $\tilde{A}(t - t_0) = \displaystyle\int_{t_0}^{t} N(s)ds$, where t_0 is fixed, we have that

$$\frac{\tilde{A}(t-t_0)}{t-t_0} = \frac{A(t)}{t-t_0} - \frac{A(t_0)}{t-t_0} \to \begin{cases} \theta - 0 \\ \lambda\mu - 0 \end{cases}$$

$$= \begin{cases} \theta & \text{for the On–Off input} \\ \lambda\mu & \text{for the Poisson input} \end{cases} \text{ as } t \to \infty.$$

Let $\tau(H) = \inf\{t \geq 0 : X(t) > H\}$ be the time until the server's buffer content reaches the level H (overflows). We are interested in the behavior of $E\tau(H)$ as $H \to \infty$. We introduce two additional random times. Let T_H be the first time a session (an On session in the On–Off case) of length at least H starts and define $\tau_2(H)$ as follows. Eliminate the leaky bucket, and let all the work go instantaneously (as opposed to in a fluid manner) to the server's buffer. Let $\tau_2(H)$ be the first time until the content of the server's buffer under the modified scenario reaches the level H.

Lemma 3.2. *For $\epsilon > 0$ small enough*

$$\lim_{H\to\infty} P(\tau_2(H) < T_{\epsilon H}) = 0.$$

Proof.

In order to prove the statement we are going to consider a simpler process, and we will prove that a buffer does not overflow before a session of length at least ϵH occurs.

Consider a modified system in which the input process results from truncating the On sessions at ϵH. That is, if an On session is longer than ϵH then we just let it be of length ϵH. On the event of interest, $\{\tau_2(H) < T_{\epsilon H}\}$, the original and modified processes coincide until time $\tau_2(H)$.

In the modified system there is no leaky bucket, and all the work goes immediately into the server's buffer. Furthermore, the way the work is added and removed from that buffer is different now. In the On–Off case, when an On session starts (say, of length l_{on} in the original process, so that it is of length $\tilde{l}_{\text{on}} = \min(l_{\text{on}}, \epsilon H)$ in the modified process) then the amount $\tilde{l}_{\text{on}}(1-r)$ is added immediately to the server's buffer. On the other hand, when an Off session starts(say, of length l_{off}) the buffer content goes down by $l_{\text{off}}r$ immediately if there is that much work left, otherwise it just goes down to zero. Similarly, in the Poisson case, when a session ends (after time $l = \min(l, \epsilon H)$) the buffer content goes down by Ir immediately if there is that much work left, otherwise it just goes down to zero, where I is the next interarrival time (note that $I \sim \exp(\lambda)$).

Clearly, under the new rules the content of the server's buffer will reach level H not later than the time $\tau_2(H)$, and we will still use the same notation, $\tau_2(H)$, to denote the time the content of that buffer reaches level H.

In the argument below the reader should mentally substitute F for F_{on} and α for α_{on} any time one considers the Poisson input case as opposed to the On–Off input case.

Now we are going to break up the probability we want to calculate into "cycles". Consider the following stopping times:

$$R_0 = 0,$$

$$R_i = \inf\left\{ t > R_{i-1} : \tilde{X}(t) = 0, \sup_{R_{i-1} \le s \le t} \tilde{X}(s) > 0 \right\}, \text{ for } i \ge 1,$$

where $\{\tilde{X}(t), t \ge 0\}$ is the modified process as described before. Then we have

$$P(\tau_2(H) < T_{\epsilon H}) = \sum_{n=1}^{\infty} P(R_{n-1} < \tau_2(H) \le R_n, \tau_2(H) < T_{\epsilon H}),$$

where $T_{\epsilon H}$ is the time we have to wait in the original system to see a session of length at least ϵH. Now, by the Strong Markov Property

$$P(R_{n-1} < \tau_2(H) \le R_n, \tau_2(H) < T_{\epsilon H}) =$$

$$= P(\tau_2(H) > R_1, R_1 < T_{\epsilon H})^{n-1} P(\tau_2(H) \le R_1, \tau_2(H) < T_{\epsilon H}).$$

So we have that

$$P(\tau_2(H) < T_{\epsilon H}) = \frac{P(\tau_2(H) \le R_1, \tau_2(H) < T_{\epsilon H})}{1 - P(\tau_2(H) > R_1, R_1 < T_{\epsilon H})} \le$$

$$\le \frac{P(\tau_2(H) \le R_1, \tau_2(H) < T_{\epsilon H})}{1 - F_{\text{on}}(\epsilon H)},$$

since $P(\tau_2(H) > R_1, R_1 < T_{\epsilon H}) \le P(R_1 < T_{\epsilon H})$ is bounded from above by the probability that the length of the first arriving session does not exceed ϵH.

In order to prove that this expression goes to zero we want to know how $P(\tau_2(H) \leq R_1, \tau_2(H) < T_{\epsilon H})$ behaves as $H \to \infty$, since we know that

$$1 - F_{\text{on}}(\epsilon H) = L_{\text{on}}(\epsilon H)(\epsilon H)^{-\alpha_{\text{on}}}, \text{ as } H \to \infty.$$

Consider a random walk defined as follows. In the On–Off case, let $(\tilde{l}^i_{\text{on}})$ and (l^i_{off}) be two independent sequences of iid On times, truncated at ϵH, as above, and Off times, accordingly. Let $Z_0 = 0$ and $Z_i = \tilde{l}^i_{\text{on}}(1-r) - l^i_{\text{off}} r$, for $i \geq 1$. In the Poisson case, let $(\tilde{l}^i)$ be an iid sequence of session lengths, truncated at ϵH as above, and (I^i) independent from it, an iid sequence of exponential random variables with parameter λ. Let $Z_0 = 0$, $Z_i = \tilde{l}^i - I^i r$, for $i \geq 1$. Then the random walk defined by these Z's, $\sum_{i=0}^n Z_i$, is a negative drift random walk, since $EZ_i < 0$. We are interested in the probability that this random walk reaches the level H (we will call that probability P_H), since

$$P(\tau_2(H) \leq R_1, \tau_2(H) < T_{\epsilon H}) \leq P(\tau_2(H) \leq R_1) =$$

$$= P(\text{the random walk of the } Z\text{'s reaches } H \text{ before reaching zero}) \leq P_H.$$

Now, in order for the random walk to reach level H, it has to first reach level ϵH. Since each step of the walk is at most ϵH, when we first reach level ϵH the walk can at most be at level $2\epsilon H$, so we have that

$$P_H = P_{\epsilon H} P(\text{the random walk reaches } H | \text{reached } \epsilon H) \leq$$

$$\leq P_{\epsilon H} P_{(1-2\epsilon)H} \leq P_{\epsilon H}^{(1-\epsilon)/2\epsilon}.$$

Then,

$$\lim_{H\to\infty} P(\tau_2(H) < T_{\epsilon H}) \leq \lim_{H\to\infty} \frac{P_{\epsilon H}^{(1-\epsilon)/2\epsilon}}{L_{\text{on}}(\epsilon H)(\epsilon H)^{-\alpha_{\text{on}}}}.$$

If the walk wasn't truncated we would have that, by Embrechts and Veraverbeke (1982), P_H is regularly varying with exponent $\alpha_{\text{on}} - 1$, and notice that truncating in our case the steps of the walk at ϵH can only make $P_{\epsilon H}$ smaller. Therefore, for $\epsilon < \frac{\alpha_{\text{on}}-1}{3\alpha_{\text{on}}-1}$

$$\lim_{H\to\infty} P(\tau_2(H) < T_{\epsilon H}) = 0.$$

Q.E.D.

Lemma 3.3. *For any* $0 < \epsilon < 1$

$$\lim_{H\to\infty} P(\tau(H) < T_{\beta(1-\epsilon)H}) = 0,$$

where

$$\beta = \begin{cases} \frac{\gamma-\theta}{(\gamma-r)(1-\theta)} & \textit{for the On–Off input.} \\ \frac{\gamma-\lambda\mu}{\gamma-r} & \textit{for the Poisson input.} \end{cases}$$

Proof.

For any $\delta > 0$ we have

$$P\left(\tau(H) < T_{\beta(1-\epsilon)H}\right) \leq$$

$$\leq P\left(\tau_2(\epsilon H/2) < T_{\delta H}\right) + P\left(T_{\delta H} \leq \tau(H) < T_{\beta(1-\epsilon)H}, \tau_2(\epsilon H/2) \geq T_{\delta H}\right).$$

Now, by Lemma 3.2, as long as δ/ϵ is small enough,

$$\lim_{H\to\infty} P\left(\tau_2(\epsilon H/2) < T_{\delta H}\right) = 0. \tag{3.6}$$

Furthermore, for $H > 0$ let V_H be the first time after time T_H that both buffers are empty and the leaky bucket is full. Then we have

$$P\left(T_{\delta H} \leq \tau(H) < T_{\beta(1-\epsilon)H}, \tau_2(\epsilon H/2) \geq T_{\delta H}\right) =$$
$$= P\left(T_{\delta H} \leq \tau(H) < T_{\beta(1-\epsilon)H}, \tau_2(\epsilon H/2) \geq T_{\delta H}, \tau(H) \leq V_{\delta H}\right) +$$
$$+P\left(T_{\delta H} \leq \tau(H) < T_{\beta(1-\epsilon)H}, \tau_2(\epsilon H/2) \geq T_{\delta H}, \tau(H) > V_{\delta H}\right) :=$$
$$:= q_1(H) + q_2(H).$$

Observe that by the Strong Markov Property

$$q_2(H) \leq P\left(T_{\delta H} \neq T_{\beta(1-\epsilon)H}, V_{\delta H} < \tau(H) \leq T_{\beta(1-\epsilon)H}\right) \leq$$
$$\leq P\left(T_{\delta H} \neq T_{\beta(1-\epsilon)H}\right) P\left(\tau(H) < T_{\beta(1-\epsilon)H}\right).$$

Therefore,

$$P\left(\tau(H) < T_{\beta(1-\epsilon)H}\right) \leq \frac{P\left(\tau_2(\epsilon H/2) < T_{\delta H}\right) + q_1(H)}{P\left(T_{\delta H} = T_{\beta(1-\epsilon)H}\right)}.$$

For every $\delta > 0$

$$\lim_{H\to\infty} P\left(T_{\delta H} = T_{\beta(1-\epsilon)H}\right) = l_\delta > 0.$$

Therefore, it follows from (3.6) that the statement of the lemma will follow once we show that for all $\delta > 0$

$$\lim_{H\to\infty} q_1(H) = 0. \tag{3.7}$$

To this end let us introduce some additional notation. Consider first the case of the On–Off input. For $H > 0$ let X_H be the length of the On session arriving at time T_H, and let W_H be the first time after time $T_H + X_H$ (end of transmission of that session) that buffer Y is empty.

We increase the probability $q_1(H)$ by moving, at time $T_{\delta H}$, the entire content of buffer Y to buffer X, and making the leaky bucket full. Note that, on the event whose probability is $q_1(H)$, this results in the content of buffer X being less than $\epsilon H/2$. We now work with the modified system (but we use the old notation). We have

$$q_1(H) \leq P\left(\tau(H) < T_{\beta(1-\epsilon)H}, \tau(H) \leq W_{\delta H}\right) +$$
$$+P\left(\tau(H) < T_{\beta(1-\epsilon)H}, W_{\delta H} < \tau(H) \leq V_{\delta H}\right) := \tag{3.8}$$
$$:= P\left(B_{11}(H)\right) + P\left(B_{12}(H)\right) := q_{11}(H) + q_{12}(H).$$

Consider first the event described by $B_{11}(H)$. Obviously, from time $T_{\delta H}$ to time $T_{\delta H} + X_{\delta H}$ the content of buffer X goes up. At the latter time the content of buffer Y is

$$Y(T_{\delta H} + X_{\delta H}) = (1-\gamma)\left(X_{\delta H} - \frac{C}{1-\gamma}\right) = (1-\gamma)X_{\delta H} - C,$$

while the content of buffer X is

$$X(T_{\delta H} + X_{\delta H}) = X(T_{\delta H}) + (1-r)\frac{C}{1-\gamma} + (\gamma - r)\left(X_{\delta H} - \frac{C}{1-\gamma}\right) =$$
$$= X(T_{\delta H}) + (\gamma - r)X_{\delta H} + C,$$

provided that $H > C/\delta$. Note that, on our event, for $H > 0$ large enough,

$$X(T_{\delta H}) + (\gamma - r)X_{\delta H} + C \leq \frac{\epsilon H}{2} + C + (\gamma - r)\beta(1-\epsilon)H \leq$$

$$\leq \frac{2\epsilon}{3}H + (\gamma - r)\frac{(1-\epsilon)(\gamma-\theta)}{(\gamma-r)(1-\theta)}H < H.$$

Hence, the content of buffer X cannot reach level H before time $T_{\delta H} + X_{\delta H}$.

Note, further, that the content of buffer X also goes up from time $T_{\delta H} + X_{\delta H}$ to time $W_{\delta H}$. Let $D_{\delta H} = W_{\delta H} - \left(T_{\delta H} + X_{\delta H}\right)$ be the length of that time interval. Notice that, on the event $B_{11}(H)$, for large $H > 0$,

$$D_{\delta H} \geq \frac{1}{\gamma}Y(T_{\delta H} + X_{\delta H}) = \frac{1-\gamma}{\gamma}X_{\delta H} - \frac{C}{\gamma} \geq \tag{3.9}$$

$$\geq \frac{1-\gamma}{\gamma}\delta H - \frac{C}{\gamma} \geq \frac{1-\gamma}{2\gamma}\delta H.$$

For $t > 0$ let $\tilde{A}(t)$ be the total amount of work brought in by the On sessions starting in the time interval $[T_{\delta H} + X_{\delta H}, T_{\delta H} + X_{\delta H} + t]$. By Lemma 3.1, for every $\rho > 0$

$$\lim_{t\to\infty} P\left(\tilde{A}(s) \geq (\theta+\rho)s \text{ for some } s \geq t\right) = 0. \tag{3.10}$$

Write, for $\rho > 0$,

$$q_{11}(H) = P\left(B_{11}(H) \cap \left\{\tilde{A}(D_{\delta H}) \geq (\theta+\rho)D_{\delta H}\right\}\right) + \tag{3.11}$$

$$+P\left(B_{11}(H) \cap \left\{\tilde{A}(D_{\delta H}) < (\theta+\rho)D_{\delta H}\right\}\right) := q_{111}(H) + q_{112}(H).$$

It follows immediately by (3.9) and (3.10) that for any $\rho > 0$

$$\lim_{H\to\infty} q_{111}(H) = 0. \tag{3.12}$$

Note, further, that

$$D_{\delta H} = \frac{Y(T_{\delta H} + X_{\delta H}) + \tilde{A}(D_{\delta H})}{\gamma}.$$

On the event whose probability $q_{112}(H)$ measures, we have, therefore,

$$D_{\delta H} \leq \frac{Y(T_{\delta H} + X_{\delta H}) + (\theta+\rho)D_{\delta H}}{\gamma},$$

and so, as long as $\rho < \gamma - \theta$,

$$D_{\delta H} \leq \frac{1}{\gamma-\theta-\rho}Y(T_{\delta H} + X_{\delta H}) \leq \frac{1-\gamma}{\gamma-\theta-\rho}X_{\delta H} \leq$$

$$\leq \frac{1-\gamma}{\gamma-\theta-\rho}\frac{(1-\epsilon)(\gamma-\theta)}{(\gamma-r)(1-\theta)}H$$

(recall that $X_{\delta H} \leq \beta(1-\epsilon)H$). Therefore, at time $W_{\delta H}$ the content of buffer X is

$$X(W_{\delta H}) = X(T_{\delta H} + X_{\delta H}) + (\gamma - r)D_{\delta H} \leq$$

$$\leq H\left(\frac{2\epsilon}{3} + (1-\epsilon)\frac{\gamma-\theta}{1-\theta} + (1-\epsilon)\frac{(1-\gamma)(\gamma-\theta)}{(\gamma-\theta-\rho)(1-\theta)}\right) = dH,$$

for some $0 < d < 1$ as long as ρ is small enough. Therefore, for all ρ small enough,

$$\lim_{H\to\infty} q_{112}(H) = 0, \tag{3.13}$$

and it follows from (3.12) and (3.13) that

$$\lim_{H\to\infty} q_{11}(H) = 0. \tag{3.14}$$

It remains, therefore, to consider the probability $q_{12}(H)$. The same decomposition as that in (3.11) shows that we can write for any $\rho > 0$

$$q_{12}(H) = P\left(B_{12}(H) \cap \left\{\tilde{A}(D_{\delta H}) \geq (\theta+\rho)D_{\delta H}\right\}\right) +$$

$$+P\left(B_{12}(H) \cap \left\{\tilde{A}(D_{\delta H}) < (\theta+\rho)D_{\delta H}\right\}\right) := q_{121}(H) + q_{122}(H).$$

We have, once again,

$$\lim_{H\to\infty} q_{121}(H) = 0, \tag{3.15}$$

by (3.9) and (3.10). Furthermore, we have already checked that, as long as ρ is small enough, we have $X(W_{\delta H}) \leq dH$ for some $0 < d < 1$. We increase the probability $q_{122}(H)$ by modifying the system as follows. At time $W_{\delta H}$ we remove the leaky bucket and buffer Y. Then the probability $q_{122}(H)$ is bounded from above by the probability that the standard system without the leaky bucket control reaches level H starting from level dH before hitting zero. Now allow buffer content to be negative (service takes place not only when there is work in the buffer, but always). Then $q_{122}(H)$ does not exceed the probability that the state of this new system ever reaches level $(1-d)H$, starting at zero. This probability however, goes to zero as $H \to \infty$ because of the negative drift. To see this simply notice that this latter probability is the same as the probability that the random walk we constructed in the proof of Lemma 3.2 ever reaches level $(1-d)H$.

Hence,

$$\lim_{H\to\infty} q_{122}(H) = 0, \tag{3.16}$$

and so

$$\lim_{H\to\infty} q_{12}(H) = 0, \tag{3.17}$$

by (3.15) and (3.16). Now (3.7) follows from (3.14) and (3.17). This finishes the argument in the case of the On–Off input.

In the case of the Poisson input, the notation is similar. For $H > 0$ let X_H be the length of the session arriving at time T_H, W_H the first time after time $T_H + X_H$ (end of transmission of that session) that buffer Y is empty, and, additionally, let R_H be the total of the remaining lengths of all sessions running just prior to time T_H.

As in the On–Off case, we increase the probability $q_1(H)$ by moving, at time $T_{\delta H}$, to buffer X the entire content of buffer Y as well as the total of the remaining lengths of all sessions running just prior to time $T_{\delta H}$ and making the leaky bucket full. Once again, on the event whose probability is $q_1(H)$, this results in the content of buffer X being less than $\epsilon H/2$, we work with the modified system and use the old notation and, finally, (3.8) is still valid. The reader will observe that the remainder of the argument below is quite similar to that above in the On–Off case, with several required modifications. Consider once again the event described by $B_{11}(H)$. For $t > 0$ let $\tilde{A}_1(t)$ be the total amount of work brought in by the sessions starting in the time interval $(T_{\delta H}, T_{\delta H} + t]$ and let $\tilde{A}_2(t)$ be the total amount of work brought in by the sessions starting in the time interval $[T_{\delta H} + X_{\delta H}, T_{\delta H} + X_{\delta H} + t]$. By the remark following Lemma 3.1, for every $\rho > 0$

$$\lim_{t\to\infty} P\left(\tilde{A}_i(s) \geq (\lambda\mu+\rho)s \text{ for some } s \geq t\right) = 0, \ i = 1, 2. \tag{3.18}$$

Write

$$q_{11}(H) = P\left(B_{11}(H) \cap \left\{\tilde{A}_1(X_{\delta H}) \geq (\lambda\mu+\rho)D_{\delta H}\right\}\right) + \tag{3.19}$$

$$+P\left(B_{11}(H) \cap \left\{\tilde{A}_1(X_{\delta H}) < (\lambda\mu+\rho)D_{\delta H}\right\}\right) :=$$

$$:= P\left(B_{111}(H)\right) + P\left(B_{112}(H)\right) := q_{111}(H) + q_{112}(H).$$

It follows immediately from (3.18) that for any $\rho > 0$

$$\lim_{H\to\infty} q_{111}(H) = 0. \tag{3.20}$$

Consider now the probability $q_{112}(H)$. We will increase this probability by moving, at time $T_{\delta H} + X_{\delta H}$, all the remaining work in sessions running at that time (these sessions must have arrived in the time interval $(T_{\delta H}, T_{\delta H} + X_{\delta H}]$) to buffer Y. On the event we are considering, the amount of work being thus deposited to buffer Y does not exceed $(\lambda\mu+\rho)X_{\delta H}$. Observe that from time $T_{\delta H}$ to time $T_{\delta H} + X_{\delta H}$ the content of buffer X goes up. At the latter time the content of buffer Y (including the work added instantaneously to it, as described above) satisfies

$$Y(T_{\delta H}+X_{\delta H}) \leq (1-\gamma)\left(X_{\delta H} - \frac{C}{1-\gamma}\right) + (\lambda\mu+\rho)X_{\delta H} = (1-\gamma+\lambda\mu+\rho)X_{\delta H} - C,$$

while the content of buffer X is

$$X(T_{\delta H} + X_{\delta H}) = X(T_{\delta H}) + (1-r)\frac{C}{1-\gamma} + (\gamma - r)\left(X_{\delta H} - \frac{C}{1-\gamma}\right) =$$

$$= X(T_{\delta H}) + (\gamma - r)X_{\delta H} + C,$$

provided that $H > C/\delta$. Note that, on our event, for $H > 0$ large enough,

$$X(T_{\delta H}) + (\gamma - r)X_{\delta H} + C \leq \frac{\epsilon H}{2} + C + (\gamma - r)\beta(1-\epsilon)H \leq$$

$$\leq \frac{2\epsilon}{3}H + (\gamma - r)\frac{(1-\epsilon)(\gamma - \lambda\mu)}{\gamma - r}H < H.$$

Hence, the content of buffer X cannot reach level H before time $T_{\delta H} + X_{\delta H}$.

Note, further, that the content of buffer X also goes up from time $T_{\delta H} + X_{\delta H}$ to time $W_{\delta H}$. Let $D_{\delta H} = W_{\delta H} - \left(T_{\delta H} + X_{\delta H}\right)$ be the length of that time interval. Notice that, on the event we are considering, for large $H > 0$,

$$D_{\delta H} \geq \frac{1}{\gamma}Y(T_{\delta H} + X_{\delta H}) = \frac{1-\gamma+}{\gamma}X_{\delta H} - \frac{C}{\gamma} \geq \tag{3.21}$$

$$\geq \frac{1-\gamma}{\gamma}\delta H - \frac{C}{\gamma} \geq \frac{1-\gamma}{2\gamma}\delta H.$$

Write, for $\rho > 0$,

$$q_{112}(H) = P\left(B_{112}(H) \cap \left\{\tilde{A}_2(D_{\delta H}) \geq (\lambda\mu+\rho)D_{\delta H}\right\}\right) + \tag{3.22}$$

$$+P\left(B_{112}(H) \cap \left\{\tilde{A}_2(D_{\delta H}) < (\lambda\mu+\rho)D_{\delta H}\right\}\right) := q_{1121}(H) + q_{1122}(H).$$

It follows immediately by (3.21) and (3.18) that for $\rho > 0$

$$\lim_{H\to\infty} q_{1121}(H) = 0. \tag{3.23}$$

Note, further, that, since at time $T_{\delta H} + X_{\delta H}$ there are no sessions present in the system,

$$D_{\delta H} = \frac{Y(T_{\delta H} + X_{\delta H}) + \tilde{A}_2(D_{\delta H})}{\gamma}.$$

On the event whose probability $q_{1122}(H)$ measures, we have, therefore,

$$D_{\delta H} \leq \frac{Y(T_{\delta H} + X_{\delta H}) + (\lambda\mu + \rho)D_{\delta H}}{\gamma},$$

and so, as long as $\rho < \gamma - \lambda\mu$,

$$D_{\delta H} \leq \frac{1}{\gamma - \lambda\mu - \rho} Y(T_{\delta H} + X_{\delta H}) \leq \frac{1 - \gamma + \lambda\mu + \rho}{\gamma - \lambda\mu - \rho} X_{\delta H} \leq$$

$$\leq \frac{1 - \gamma + \lambda\mu + \rho}{\gamma - \lambda\mu - \rho} \frac{(1-\epsilon)(\gamma - \lambda\mu)}{\gamma - r} H$$

(recall that $X_{\delta H} \leq \beta(1-\epsilon)H$). Therefore, at time $W_{\delta H}$ the content of buffer X is

$$X(W_{\delta H}) = X(T_{\delta H} + X_{\delta H}) + (\gamma - r)D_{\delta H} \leq$$

$$\leq H\left(\frac{2\epsilon}{3} + (1-\epsilon)(\gamma - \lambda\mu) + (1-\epsilon)\frac{(1 - \gamma + \lambda\mu + \rho)(\gamma - \lambda\mu)}{\gamma - \lambda\mu - \rho}\right) = dH,$$

for some $0 < d < 1$ as long as ρ is small enough. Therefore, for all ρ small enough,

$$\lim_{H\to\infty} q_{1122}(H) = 0, \tag{3.24}$$

and it follows from (3.20), (3.23) and (3.24) that

$$\lim_{H\to\infty} q_{11}(H) = 0. \tag{3.25}$$

It remains, therefore, to consider the probability $q_{12}(H)$. But using the same arguments as for the On–Off case we get that

$$\lim_{H\to\infty} q_{12}(H) = 0. \tag{3.26}$$

Now (3.17) follows from (3.25) and (3.26).

Q.E.D.

We are ready now to prove the main result.

Proof of Theorem 3.1

For the upper bound it is enough to show one way in which the buffer can overflow. Now, in order to get a sharp upper bound we want to consider the most likely scenario in which the buffer will overflow. We will prove that in this case, in the presence of heavy tails and having $\gamma < 1$, what usually causes the buffer to overflow is one single very long session.

One overflow scenario could be the following. Consider a long session of size S that arrives at the system at a renewal time. We can assume that the session arrives at a renewal time since if it doesn't then the buffer will overflow even sooner, and we are just looking at an upper bound. As it turns out , this will not matter even for the lower bound, since asymptotically what is going to matter is how long it takes for this long session to arrive.

We start with a heuristic calculation of just how long this long session of size S has to be in order to cause buffer overflow. Consider, for the moment, a long session of size S arriving at time zero. Then $Z\left(\frac{C}{1-\gamma}\right) = 0$, since $Z(\cdot)$ decreases at rate $1-\gamma$, and $X\left(\frac{C}{1-\gamma}\right) = \frac{(1-r)C}{1-\gamma}$, since $X(\cdot)$ increases at rate $1-r$ when $Z(t)$

is positive. After time $\frac{C}{1-\gamma}$, $X(\cdot)$ grows linearly at rate $\gamma - r$ for as long as the buffer of the leaky bucket is not empty. Since we are interested in the result when $H \to \infty$, we can safely assume that $\frac{(1-r)C}{1-\gamma} < H$.

If $X(\cdot)$ continues to grow linearly at the same rate, which will happen if S is large enough (we will see in a moment just how large it has to be), then the time it takes from the beginning of the long session until the buffer overflows, x_0, satisfies the relation

$$H - \frac{C(1-r)}{1-\gamma} = (\gamma - r)\left(x_0 - \frac{C}{1-\gamma}\right) \Longrightarrow x_0 = \frac{H-C}{\gamma - r},$$

which results from letting $X(\cdot)$ grow at rate $1-r$ until time $\frac{C}{1-\gamma}$ and then grow at rate $\gamma - r$ until the buffer reaches level H.

Now we need to calculate how large S has to be in order for the scenario we just described to happen. In the On–Off input case, observe that after this long session ends the amount of work coming into the system until time x_0 is about the expected amount of work, that is θ times the length of the time interval. In particular, for any $0 < \epsilon < \theta$, the amount of work coming into the system since the long On session ends during a long time interval is unlikely to be less than $(\theta - \epsilon)$ times the length of the interval (this is, below the mean expected amount of work for that period of time) and in that case the buffer of the leaky bucket will not become empty until $Y(S)/(\gamma - \theta + \epsilon)$ units of time later, during which time the content of buffer X continues to grow at rate $\gamma - r$. (Note that for this to happen it is enough that the amount of work coming into the system in $Y(S)/\gamma$ or more units of time after the end of the long session is at least $(\theta - \epsilon)$ times the length of the interval, since the buffer of the leaky bucket will not become empty before that.) Since

$$Y(S) = (1-\gamma)\left(S - \frac{C}{1-\gamma}\right), \tag{3.27}$$

then the minimum length of session so that the buffer overflows must satisfy

$$\frac{(1-\gamma)\left(S - \frac{C}{1-\gamma}\right)}{(\gamma - \theta + \epsilon)} \geq \left(\frac{H-C}{\gamma - r} - S\right).$$

Figure 2 provides a graphical description of the above discussion. Solving for S implies that the minimum length of session so that the buffer overflows, $S(H)$, is

$$S(H) = \frac{\gamma - \theta + \epsilon}{(\gamma - r)(1 - \theta + \epsilon)} H + \frac{\theta - r - \epsilon}{(\gamma - r)(1 - \theta + \epsilon)} C. \tag{3.28}$$

That is, we expect that if a session of the size $S \geq S(H)$ arrives at time zero, the buffer will overflow no later than at time x_0.

To calculate the corresponding required session length in the Poisson case we observe that after time $\frac{C}{1-\gamma}$ the amount of work coming into the system until the buffer overflows is about the expected amount of work, that is $\lambda\mu$ times the length of the time interval. In particular, for any $0 < \epsilon < \lambda\mu$, if the amount of work coming into the system after time $\frac{C}{1-\gamma}$ during a long time interval is unlikely to be less than $(\lambda\mu - \epsilon)$ times the length of the interval (this is below the mean expected amount of work for that period of time) and in that case the buffer of the leaky bucket will not become empty until $Y(S)/(\gamma - \lambda\mu + \epsilon)$ units of time after the long

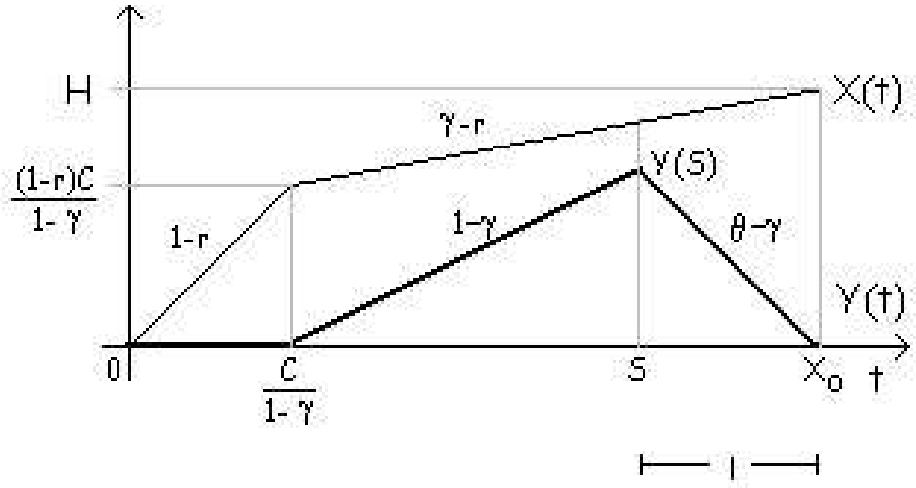

FIGURE 2. Shortest session S that makes a buffer of size $H >> 0$ overflow in the On–Off case.

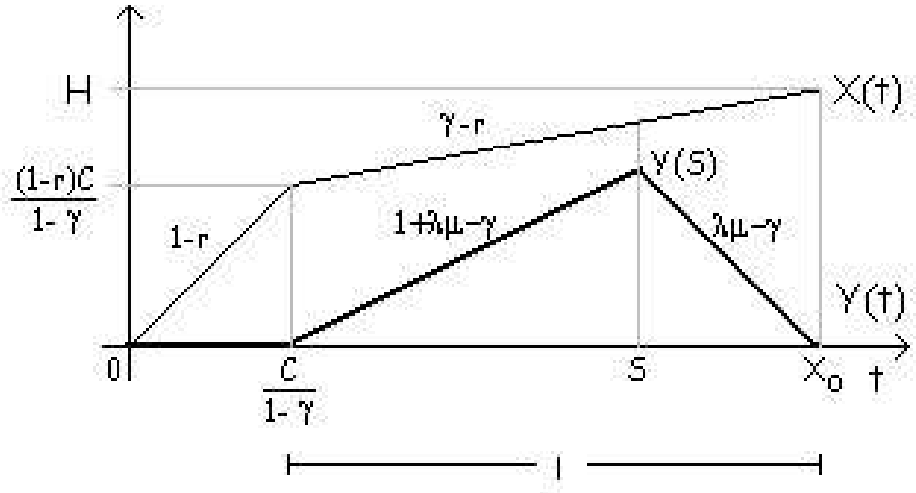

FIGURE 3. Shortest session S that makes a buffer of size $H >> 0$ overflow in the Poisson case.

session ends, during which time the content of buffer X continues to grow at rate $\gamma - r$.

Since we have additionally, under this scenario,

$$Y(S) \geq (1 + \lambda\mu - \gamma - \epsilon)\left(S - \frac{C}{1-\gamma}\right),$$

then it is enough for the minimum length of session to satisfy (see figure 3)

$$\frac{(1+\lambda\mu-\gamma-\epsilon)\left(S-\frac{C}{1-\gamma}\right)}{(\gamma-\lambda\mu+\epsilon)} \geq \left(\frac{H-C}{\gamma-r}-S\right).$$

See Figure 3 for graphical presentation of the above argument.

Solving for S implies that a session of the length $S(H)$ given by

$$S(H) = \frac{\gamma-\lambda\mu+\epsilon}{\gamma-r}H + \frac{(\lambda\mu-\epsilon-r(1+\lambda\mu-\gamma-\epsilon))}{(\gamma-r)(1-\gamma)}C \tag{3.29}$$

will cause the buffer X to overflow. That is, we expect that a session of size $S \geq S(H)$ given in (3.29) arriving at time zero will cause the buffer to overflow no later than at time x_0.

Clearly, a session as long as that given by either (3.28) or (3.29) does not arrive at time zero, so we need to know how long we have to wait for such a long session to occur. In the On–Off case, we know that

$$P(l_{\text{on}} > S(H)) \sim S(H)^{-\alpha_{\text{on}}} L_{\text{on}}(S(H)),$$

so we need to wait for approximately

$$\frac{1}{P(l_{\text{on}} > S(H))} = \frac{S(H)^{\alpha_{\text{on}}}}{L_{\text{on}}(S(H))}$$

On sessions for this to happen. We need now to calculate how long we have to wait to get that many On sessions. Since each On session is followed by an Off session, we expect to have to wait for $\left(\frac{S(H)^{\alpha_{\text{on}}}}{L_{\text{on}}(S(H))}-1\right)(\mu_{\text{on}}+\mu_{\text{off}})$ units of time for $\frac{S(H)^{\alpha_{\text{on}}}}{L_{\text{on}}(S(H))}$ On sessions. Then, if H is big enough and if after the end of the long session the work arrives to the system at the rate of at least $(\theta-\epsilon)$, then we do not expect the time until overflow to exceed

$$\frac{S(H)^{\alpha_{\text{on}}}}{L_{\text{on}}(S(H))}(\mu_{\text{on}}+\mu_{\text{off}}) + x_0. \tag{3.30}$$

An identical reasoning shows that in the Poisson input case we do not expect the time until overflow to exceed

$$\frac{S(H)^{\alpha}}{\lambda L(S(H))} + x_0. \tag{3.31}$$

Finally, if, in the On–Off input case, the rate at which the work arrives to the system after the end of the long session over the relevant time interval happens to be less than $(\theta-\epsilon)$ (this is the minimal rate we expected), or if the rate at which work arrives to the system after time $\frac{C}{1-\gamma}$ over the relevant interval happens to be less than $(\lambda\mu-\epsilon)$ (this is, once again, the minimal rate we expected), we just empty the system at that time and wait again for a long On session to come.

Now we provide a rigorous argument. In the On–Off input case, let

$$l(H) = S(H) - \frac{C}{1-\gamma},$$

with $S(H)$ given by (3.28). Note that $l(H) \sim \text{const}H \to \infty$ as $H \to \infty$. Let $p_H := P\big(A(l) \geq (\theta-\epsilon)l \text{ for all } l \geq l(H)\big)$. By Lemma 3.1 we have that $p_H \to 1$ as

$H \to \infty$. Let us agree to call any session whose length is at least $S(H)$ in (3.28) a sufficiently long session. Let

$$r(H) = \frac{H - C}{\gamma - r} - S(H)$$

be the time after the end of a shortest possible sufficiently long session after which the server buffer is guaranteed to overflow if the work keeps arriving at the minimal rate we expect. Let W_k and V_k be, correspondingly, the arrival time and the end time of the kth sufficiently long session, $k \geq 1$, and $V_0 = 0$. Define events

$$B_k = \{W_k - V_{k-1} > r(H) \text{ and the amount of work arriving in } s \text{ units of time after the end of the } k\text{th sufficiently long session is at least } (\theta - \epsilon)s \text{ for every } l(H) \leq s \leq r(H)\}, \text{ for } k \geq 1.$$

Note that these events are independent, and

$$P(B_k) \geq P(W_1 > r(H))p_H, \text{ for all } k.$$

Since $p_H \to 1$ and $P(W_1 > r(H)) \to 1$ as $H \to \infty$, we see that $P(B_1) \to 1$ as $H \to \infty$.

The random variable N defined to be equal to k on $B_k \cap \left(\cap_{j=1}^{k-1} B_j^c\right)$ (and infinite outside of the union of these events) is a.s. finitely valued, and $\tau(H) \leq W_N + x_0$. Therefore,

$$E\tau(H) \leq x_0 + EW_N. \tag{3.32}$$

Similarly, in the Poisson input case, we let $p_H := P(A(l) \geq (\lambda\mu - \epsilon)l, \text{ for all } l \geq S(H))$, where A is the total input process defined in Lemma 3.1 and, as before, by Lemma 3.1 we have that $p_H \to 1$ as $H \to \infty$. Let us agree to call any session whose length is at least $S(H)$ in (3.29) a sufficiently long session. Let W_k be the arrival time of the kth sufficiently long session, $k \geq 1$, and $W_0 = 0$. Define events

$$B_k = \{W_k - W_{k-1} > S(H) \text{ and the amount of work arriving in } s \text{ units of time after the arrival of the } k\text{th sufficiently long session is at least } (\lambda\mu - \epsilon)s \text{ for every } S(H) \leq s \leq \frac{H - C}{\gamma - r}\}, \text{ for } k \geq 1.$$

Note that these events are independent, and

$$P(B_k) \geq P(W_1 > S(H))p_H, \text{ for all } k.$$

Since $p_H \to 1$ and $P(W_1 > S(H)) \to 1$ as $H \to \infty$, we see that $P(B_1) \to 1$ as $H \to \infty$.

As before, we define a random variable N to be equal to k on $B_k \cap \left(\cap_{j=1}^{k-1} B_j^c\right)$ (and infinite outside of the union of these events) is a.s. finitely valued, and $\tau(H) \leq W_N + x_0$. In particular, (3.32) still holds.

In both On–Off and Poisson input cases,write

$$EW_N = E\left(\sum_{j=k}^{N}(W_k - W_{k-1})\right) = EW_1 + \sum_{k=2}^{\infty} E\left((W_k - W_{k-1})1\!\!1_{N \geq k}\right) \leq$$

$$\leq EW_1 + (E(W_1^p))^{1/p} \sum_{k=2}^{\infty} \left(P(N \geq k)\right)^{1/q},$$

(choose $1 < p < \alpha$ as in the statement of the theorem in the On–Off input case) and $1/p + 1/q = 1$. Note that

$$EW_1 \leq \begin{cases} \frac{S(H)^{\alpha_{\rm on}}}{L_{\rm on}(S(H))}(\mu_{\rm on} + \mu_{\rm off}) & \text{for the On–Off input} \\ \frac{S(H)^{\alpha}}{\lambda L(S(H))} & \text{for the Poisson input} \end{cases}$$

and we will check below that

$$(E(W_1^p))^{1/p} \leq {\rm O}\Big(EW_1\Big) \quad \text{as } H \to \infty. \tag{3.33}$$

Finally,

$$P(N \geq k) = P\left(\bigcap_{j=1}^{k-1} B_j^c\right) = (1 - P(B_1))^{k-1},$$

which implies that $EW_N \sim EW_1$ as $H \to \infty$, and so in the On–Off input case,

$$\limsup_{H\to\infty} \left(\frac{S(H)^{\alpha_{\rm on}}}{L_{\rm on}(S(H))}(\mu_{\rm on} + \mu_{\rm off})\right)^{-1} EW_N \leq 1.$$

Using (3.32), the expression (3.28) for $S(H)$ and the fact that x_0 grows linearly fast with H, we conclude that

$$\limsup_{H\to\infty} H^{-\alpha_{\rm on}} L_{\rm on}(H) E\tau(H) \leq (\mu_{\rm on} + \mu_{\rm off}) \left(\frac{\gamma - \theta + \epsilon}{(\gamma - r)(1 - \theta + \epsilon)}\right)^{\alpha_{\rm on}}.$$

Letting $\epsilon \to 0$ we obtain the upper bound

$$\limsup_{H\to\infty} H^{-\alpha_{\rm on}} L_{\rm on}(H) E\tau(H) \leq (\mu_{\rm on} + \mu_{\rm off}) \left(\frac{\gamma - \theta}{(\gamma - r)(1 - \theta)}\right)^{\alpha_{\rm on}}. \tag{3.34}$$

An identical argument gives

$$\limsup_{H\to\infty} H^{-\alpha} L(H) E\tau(H) \leq \frac{1}{\lambda}\left(\frac{\gamma - \lambda\mu}{\gamma - r}\right)^{\alpha} \tag{3.35}$$

in the Poisson input case. For the lower bound note that for $1 < p < \alpha$ the p-th moment exists, both for the On and Off session length distributions in the On–Off input case and for the session length distribution in the Poisson input case. In either case, for any $0 < \epsilon < 1$

$$E\tau(H) \geq E(\tau(H) 1\!\!1_{(\tau(H) \geq T_{\beta(1-\epsilon)H})}) \geq E(T_{\beta(1-\epsilon)H} 1\!\!1_{(\tau(H) \geq T_{\beta(1-\epsilon)H})}) =$$

$$= E(T_{\beta(1-\epsilon)H}) - E(T_{\beta(1-\epsilon)H} 1\!\!1_{(\tau(H) < T_{\beta(1-\epsilon)H})}),$$

where $\beta = \frac{\gamma-\theta}{(\gamma-r)(1-\theta)}$ in the On–Off input case and $\beta = \frac{\gamma-\lambda\mu}{\gamma-r}$ in the Poisson input case. Using Hölder's inequality we get that

$$E(T_{\beta(1-\epsilon)H} 1\!\!1_{(\tau(H) < T_{\beta(1-\epsilon)H})}) \leq (E(T_{\beta(1-\epsilon)H})^p)^{1/p} (E(1\!\!1^q_{(\tau(H) < T_{\beta(1-\epsilon)H})}))^{1/q} =$$

$$= (E(T_{\beta(1-\epsilon)H})^p)^{1/p} (P(\tau(H) < T_{\beta(1-\epsilon)H}))^{1/q},$$

where, as before, $1/p + 1/q = 1$.

By Lemma 3.3 we know that $\lim_{H\to\infty} P(\tau(H) < T_{\beta(1-\epsilon)H}) = 0$, so now, if we prove that $(E(T_{\beta(1-\epsilon)H})^p)^{1/p}$ is of the same order as $ET_{\beta(1-\epsilon)H}$ as $H \to \infty$, then we have proved (3.33) and that

$$\liminf_{H\to\infty} \frac{E\tau(H)}{ET_{\beta(1-\epsilon)H}} \geq 1.$$

In the On–Off input case note that

$$ET_H \geq \left(\frac{H^{\alpha_{\rm on}}}{L_{\rm on}(H)} - 1 \right) (\mu_{\rm on} + \mu_{\rm off}).$$

Therefore, we would conclude that

$$\liminf_{H\to\infty} H^{-\alpha_{\rm on}} L_{\rm on}(H) E\tau(H) \geq (1-\epsilon)^{\alpha_{\rm on}} (\mu_{\rm on} + \mu_{\rm off}) \left(\frac{\gamma - \theta}{(\gamma - r)(1-\theta)} \right)^{\alpha_{\rm on}}.$$

And since this is true for all $0 < \epsilon < 1$ we would get that

$$\liminf_{H\to\infty} H^{-\alpha_{\rm on}} L_{\rm on}(H) E\tau(H) \geq (\mu_{\rm on} + \mu_{\rm off}) \left(\frac{\gamma - \theta}{(\gamma - r)(1-\theta)} \right)^{\alpha_{\rm on}},$$

and this would give us the lower bound matching the upper bound in (3.34), and so complete the proof of the theorem. An identical argument in the Poisson input case would give us

$$\liminf_{H\to\infty} H^{-\alpha} L(H) E\tau(H) \geq \frac{1}{\lambda} \left(\frac{\gamma - \lambda\mu}{\gamma - r} \right)^{\alpha},$$

and so give us, once again, the lower bound matching the upper bound in (3.35), and so complete the proof of the theorem in that case as well.

So we are interested in estimating $E(T_{\beta(1-\epsilon)H})^p$. In the Poisson input case observe that $T_{\beta(1-\epsilon)H}$ is an exponential random variable with mean $\lambda^{-1} P(S_1 > \beta(1-\epsilon)H)$. Therefore,

$$E(T_{\beta(1-\epsilon)H})^p = \frac{\lambda^p \Gamma(p+1)}{(P(S_1 > \beta(1-\epsilon)H))^p},$$

which is of the right order. In the On–Off input case we can write

$$T_{\beta(1-\epsilon)H} = \sum_{j=1}^{M} Z_j,$$

where the Z_j's are iid and consist of the sum of two independent random variables: one that is drawn from the Off distribution and the other that is drawn from the On distribution conditioning on it being less than $\beta(1-\epsilon)H$. So EZ_1^p is uniformly bounded from above by a constant independent of H. On the other hand M is a geometric random variable, independent of the Z_j's, with success probability $1 - F_{\rm on}(\beta(1-\epsilon)H)$. By Hölder's inequality we have that

$$T_{\beta(1-\epsilon)H} \leq \left(\sum_{j=1}^{M} Z_j^p \right)^{1/p} \left(\sum_{j=1}^{M} 1^q \right)^{1/q} \Longrightarrow$$

$$(T_{\beta(1-\epsilon)H})^p \leq \left(\sum_{j=1}^{M} Z_j^p \right) M^{p/q} \Longrightarrow$$

$$E(T_{\beta(1-\epsilon)H})^p \leq E\left(E\left(\left(\sum_{j=1}^{M} Z_j^p \right) M^{p/q} | M \right) \right) =$$

$$= E\left(E(Z_1^p) M^{1+p/q} \right) = E(Z_1^p) E(M^p).$$

Since $E(Z_1^p)$ is uniformly bounded from above, all we need to prove is that $E(M^p)$ is of the order of $(EM)^p$ as $H \to \infty$. If we let $\rho = 1 - F_{\text{on}}(\beta(1-\epsilon)H)$ then

$$E(M^p) = \sum_{k=0}^{\infty} \rho(1-\rho)^k k^p \leq \rho \sum_{k=0}^{\infty} \int_k^{k+1} (1-\rho)^{x-1} x^p dx =$$

$$= \frac{\rho}{1-\rho} \int_0^{\infty} (1-\rho)^x x^p dx = \frac{\rho}{1-\rho} \int_0^{\infty} x^p e^{-x \ln\left(\frac{1}{1-\rho}\right)} dx =$$

$$= \frac{\rho \Gamma(p+1)}{(1-\rho)\left(\ln \frac{1}{1-\rho}\right)^{p+1}} \sim \rho^{-p} \Gamma(p+1).$$

Therefore, we have the desired result in all cases.

Q.E.D.

References

A. Berger and W. Whitt (1992a): The Brownian approximation for rate-control throttles and the $G/G/1/C$ queue. *Discrete Event Dynamic Systems: Theory and Applications* 2:7–60.

A. Berger and W. Whitt (1992b): Comparison of multi-server queues with finite waiting rooms. *Stochastic Models* 8:719–732.

A. Berger and W. Whitt (1992c): The impact of a job buffer in a token-bank rate-control throttle. *Stochastic Models* 8:685–717.

A. Berger and W. Whitt (1994): The pros and cons of a job buffer in a token-bank rate-control throttle. *IEEE Transactions on Communications* 42:857–861.

O. Boxma and V. Dumas (1998): Fluid queues with long–tailed activity period distributions. *Computer Communications* 21:1509–1529. Special issue on "Stochastic Analysis and Optimization of Communication Systems".

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

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

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

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

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

P. Jelenković and A. Lazar (1999): Asymptotic results for multiplexing subexponential on–off sources. *Advances in Applied Probability* 31:394–421.

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

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

S. Resnick and G. Samorodnitsky (1999): Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues. *QUESTA* 33:43–71.

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

A. Zwart, B. Zwart and M. Mandjes (2000): Exact asymptotics for fluid queues fed by multiple heavy-tailed on-off flows. Preprint.

Department of Statistical Science, Cornell University, Ithaca, NY 14853
E-mail address: bpg5@cornell.edu

School of Operations Research and Industrial Engineering, and Department of Statistical Science, Cornell University, Ithaca, NY 14853
E-mail address: gennady@orie.cornell.edu