Информационные процессы, Том 2, № 2, 2002, стр. 186–189.
© 2002 González-Arévalo, Samorodnitsky.

Performance of the Leaky Bucket System with Long–range Dependent Input Traffic

B. Gonzalez-Arevalo and G. Samorodnitsky

Cornell University
email: bpg5@cornell.edu, gennady@orie.cornell.edu

Received October 14, 2002

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)Leland, Taqqu, Willinger and Wilson], [Paxson and Floyd(1994)], [Cunha et al.(1995)Cunha, Bestavros and Crovella], [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.

In this study 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. The 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. Let us consider two types of input processes: an On–Off process and a $M/G/\infty$ type process. Recently there has been a lot of work concerning fluid models fed by On–Off or $M/G/\infty$ type processes (see, for example, [Heath et al.(1997)Heath, Resnick and Samorodnitsky, Heath et al.(1999)Heath, Resnick and Samorodnitsky], [Jelenkovic and Lazar(1999)], a survey in [Boxma and Dumas(1998)] and a recent study in [Zwart et al.(2000)Zwart, Borst and Mandjes]). Unlike these studies 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 ([Berger and Whitt(1992a), Berger and Whitt(1992b), Berger and Whitt(1992c), Berger and Whitt(1994)]). However, to the best of our knowledge only the paper of [Vamvakos and Anantharam(1998)] looked at how the leaky bucket input control performs in the presence of a long–range dependent input. However, while [Vamvakos and Anantharam(1998)] concentrated only on the rate of decay of correlations, we look directly at system performance, specifically at the time until overflow of a large buffer. We show that in the presence of long–range dependence the buffer still overflows much more often than 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. It is important to mention that, unlike the previous authors, who looked at discrete time systems, here we investigate a fluid–type, continuous time system.

Let work arrive to the system according to some input process. We are going to consider two types of input processes: an On–Off process and a $M/G/\infty$ 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 fixed 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 buffer of the leaky bucket. The departure of work from this buffer is controlled by *tokens* that arrive at a buffer of size C at fixed 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.

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 $\alpha_{\text{on}} > 1$. 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) and hence makes it possible for the system to be stable if the service rate r is high enough. In the $M/G/\infty$ 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 $M/G/\infty$ type case we assume from now on that

$$0 < \lambda\mu < r < \gamma < 1$$

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

$$0 < \theta < r < \gamma < 1.$$

We show that under certain assumptions the rate at which the expected time until buffer overflow grows is, roughly, $1/\big(1 - F_{\text{on}}(H)\big)$ in the case of the On–Off input and $1/\big(1 - F(H)\big)$ in the case of the $M/G/\infty$ type input, where $H >> 0$ is the size of the buffer.

Let $\tau(H) = \inf\{t \geq 0 : X(t) > H\}$ be as the time until the server's buffer content reaches level H (overflows) and $\tau_Y(H) = \inf\{t \geq 0 : Y(t) > bH\}$ be the time until the leaky bucket's buffer reaches level bH. We are interested in the behavior of $E\tau(H)$ and $E\tau_Y(H)$ as $H \to \infty$. Let the size of the leaky bucket and the leaky bucket's buffer ($Y_{\max}$) grow to infinity as we let the size of the server's buffer grow to infinity in the following manner.

$$C(H) \sim aH, \text{ as } H \to \infty \text{ and}$$

$$Y_{\max}(H) \sim bH, \text{ as } H \to \infty,$$

where a and b are constants. Then we have the following results, following an argument that is based on a large deviations approach.

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

$$\lim_{H\to\infty} H^{-\alpha_{on}} L_{on}(H) E\tau(H) =$$

$$= (\mu_{on} + \mu_{off}) \left(\max\left\{ \frac{1}{1-r}, \frac{1-a}{\gamma - r} - \frac{b}{\gamma - \theta}, \frac{\gamma - \theta - (r-\theta)a}{(\gamma - r)(1-\theta)} \right\} \right)^{\alpha_{on}}. \tag{1}$$

For the $M/G/\infty$ type input process,

$$\lim_{H\to\infty} H^{-\alpha} L(H) E\tau(H) =$$

$$= \frac{1}{\lambda} \left(\max\left\{ \frac{1}{1 + \lambda\mu - r}, \frac{1-a}{\gamma - r} - \frac{b}{\gamma - \lambda\mu}, \frac{\gamma - \lambda\mu - (r - \lambda\mu)a}{\gamma - r} \right\} \right)^{\alpha}. \tag{2}$$

Remark 1. The results of Theorem 1 should be compared to the corresponding performance results without the leaky bucket input control. Then

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

for the On–Off input process (see Theorem 2.3 of [Heath et al.(1997)Heath, Resnick and Samorodnitsky]) 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 $M/G/\infty$ type 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.

For the leaky bucket's buffer we have a similar result.

Theorem 2. *For the On–Off input process, if for some $p > 1$ the p-th moment exists for the Off session length distribution, then*

$$\lim_{H\to\infty} H^{-\alpha_{on}} L_{on}(H) E\tau_Y(H) = (\mu_{on} + \mu_{off}) \left(\frac{a+b}{1-\gamma}\right)^{\alpha_{on}}. \tag{3}$$

For the $M/G/\infty$ type input process,

$$\lim_{H\to\infty} H^{-\alpha} L(H) E\tau_Y(H) = \frac{1}{\lambda}\left(\frac{a+b}{1+\lambda\mu-\gamma}\right)^{\alpha}. \tag{4}$$

With the previous results it is now easy to prove the following theorem.

Theorem 3. *For the On–Off and $M/G/\infty$ type input processes, if for some $p > 1$ the p-th moment exists for the Off session length distribution, then we have that*

$$\frac{\tau(H)}{E\tau(H)} \sim \exp(1), \textit{ as } H \to \infty.$$

REFERENCES

Berger and Whitt(1992a). A. W. Berger and W. Whitt (1992a): The Brownian Approximation for Rate-Control Throttles and the G/G/1/C Queue. *Discret Event Dynamic Systems: Theory and Applications* 2:7–60.

Berger and Whitt(1992b). A. W. Berger and W. Whitt (1992b): Comparisons of multi-server queues with finite waiting rooms. *Commun. Statist. - Stochastic Models* 8(4):719–732.

Berger and Whitt(1992c). A. W. Berger and W. Whitt (1992c): The impact of a job buffer in a token-bank rate-control throttle. *Commun. Statist. - Stochastic Models* 8(4):685–717.

Berger and Whitt(1994). A. W. 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.

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

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

Cunha et al.(1995)Cunha, Bestavros and Crovella. 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.

Heath et al.(1997)Heath, Resnick and Samorodnitsky. 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.

Heath et al.(1999)Heath, Resnick and Samorodnitsky. 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.

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

Leland et al.(1994)Leland, Taqqu, Willinger and Wilson. 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.

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

Resnick and Samorodnitsky(1999). 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.

Vamvakos and Anantharam(1998). 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.

Zwart et al.(2000)Zwart, Borst and Mandjes. A. ZWART, S. BORST and M. MANDJES (2000): Exact asymptotics for fluid queues fed by multiple heavy-tailed on-off flows. Preprint.