The Lightening Effect of Adaptive Window Control

Arzad A. Kherani and Anurag Kumar, *Senior Member, IEEE*

***Abstract*—We consider adaptive window (e.g., TCP) controlled transfer of http-like traffic (i.e., finite volume file transfers starting at random time instants) over a link, and develop an analysis for obtaining the stationary distribution of the link buffer occupancy.**

The significance of this work is that it provides a framework for analyzing the "closed-loop" behavior of the link buffer under adaptive window controlled transfer of files with a general file size distribution. We find that the tail of the stationary distribution of the buffer occupancy is *lighter* than may be expected from an open loop analysis. We also find that the tail of the buffer distribution is sensitive to the distribution of the file sizes, which provides insight into the sensitivity of TCP performance to file size distribution.

***Index Terms*—Closed-loop analysis of Internet buffers, long range dependence.**

I. Introduction

IT WAS observed in [1] that traffic processes in the Internet are long range dependent (LRD). In [2] this phenomenon was traced to heavy-tailed distribution of the file transfer volumes. Reference [3] shows that the stationary distribution of a queue fed with LRD traffic has a non-exponential tail; for example, it has been shown that an arrival rate process autocovariance that is $O(1/\tau^{\alpha-1})$, $1 < \alpha < 2$, leads to a stationary distribution of buffer occupancy that has a tail that is $O(1/x^{\alpha-1})$. These observations have been taken to indicate that the occupancy distribution in Internet router buffers have heavy tails. These observations are, however, based on an "open-loop" analysis of an LRD traffic source feeding a buffer. It has also been noted recently [4] that an understanding of traffic and buffer processes in the Internet should take into account the closed loop nature of Internet congestion control, namely TCP which is an adaptive window protocol (AWP). In this letter, we carry out such an analysis for a particular network scenario and for a general AWP.

We consider the scenario shown in Fig. 1, where a (bottleneck) link connects clients on one side to servers on the other side. The clients generate file transfer requests and the servers send the requested (finite volume) files using an AWP. We assume that the servers and clients are connected to the bottleneck link by very high speed access links. The figure also shows the Internet link's buffer containing data from ongoing file transfers; note that the number of ongoing transfers varies with time.

A. System and Traffic Assumptions

i) The end-to-end propagation delay is negligible (small so that the bandwidth delay product is much less than one packet.

Manuscript received November 15, 2003. The associate editor coordinating the review of this letter and approving it for publication was Prof. D. Petr.

The authors are with the Department of Electrical Communication Engineering Indian Institute of Science, Bangalore, 560 012, India (e-mail: alam@ece.iisc.ernet.in; anurag@ece.iisc.ernet.in).

Digital Object Identifier 10.1109/LCOMM.2003.812169

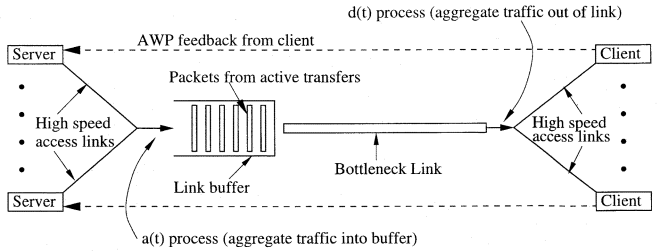

Fig. 1. AWP-controlled file transfers over a link connecting servers and clients.

ii) The link buffer implements a per flow round-robin scheduling discipline, for example, Deficit Round Robin (DRR, see [5]) or weighted fair queueing (WFQ).

iii) We assume that the link buffer is large enough so that no packet loss takes place. (Note that since the file sizes are finite the window growth is solely governed by the AWP; the window of each transfer remains finite since the volume of the transfer is finite. We wish to study the tail behavior of the stationary contents of the buffer; such an analysis would provide some insight into the loss behavior with small buffers.).

iv) Each request is for the transfer of a single file, and the files have independent and identically distributed sizes, with a general distribution.

v) The starting instants of the file transfers constitute a Poisson process of rate λ (see [4]). The instants at which new user sessions start is now accepted to be well modeled by a Poisson process (see [4]).

B. The System Model

The contents of the link buffer comprise the windows of each of the active flows; since the propagation delay is zero, the entire window of an active flow is in the link buffer. These windows are served in a round-robin manner as per (ii) of the system assumptions. For a small service quantum the round-robin discipline is simpler to study via the PS model, hence for packet size that is small compared to file sizes we can approximate the service of the windows in the link buffer by a PS discipline. Effectively we are considering the files as being composed of infinitely divisible fluid. Based on this approximation, Fig. 2 depicts the queueing equivalent of the scenario shown in Fig. 1; note that the link buffer has been replaced by a PS server. Owing to zero link propagation delay, each active flow has positive amount of outstanding data (window) in the link buffer and since these *windows* are served in a PS fashion, *it follows that the ongoing file transfers (as a whole) also get service in a PS manner irrespective of the AWP used.*

At any time instant t, an active file transfer would have successfully transferred some data to its client, some of its data would be in the link buffer (which would be the current window size of AWP controlling the transfer of the file), and the remaining data would be in the server waiting to be transferred. At any time instant we use

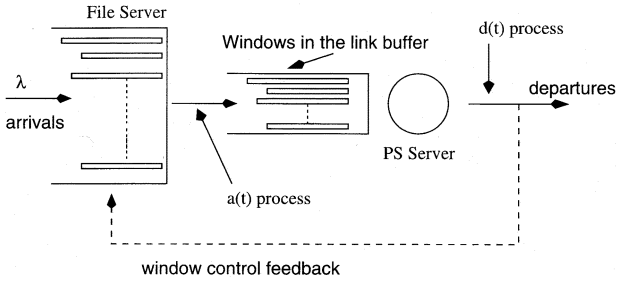

Fig. 2. Queueing equivalent of the network scenario of Fig. 1.

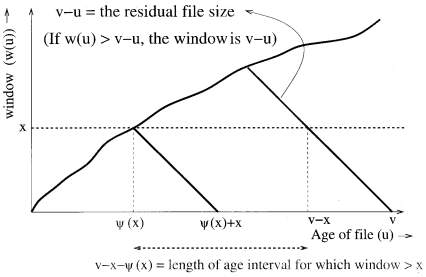

Fig. 3. An example of the evolution of the window $(w(u))$ with the age of file (u). The file of size $v \geq \psi(x) + x$ will have a window of at least x iff $u \in (\psi(x), v - x)$.

the term "age" for the amount of acknowledged data of a file. As data from a file is served (along with other files, in a PS manner) more data is pulled into the link buffer from the file server, so as to replenish the windows and to account for any window growth. Eventually the amount of data of the file in the server reduces to zero, and the entire residual file is in the link buffer.

With this model, we wish to analyze the stationary distribution of the contents of the link buffer. Call TCP-CA (TCP-SS) the TCP algorithm with initial slow start threshold set to unity (infinity). Note that if no loss occurs then TCP-SS is always in the slow start phase and TCP-CA is always in the congestion avoidance phase.

Notation: Let V denote the random variable for the file sizes brought by sessions; EV is its expectation, $V(\cdot)$ denotes its cumulative distribution function (cdf), $V^c(\cdot)$ its complementary cdf, and $V_e(\cdot)$ and $V_s(\cdot)$ denote, respectively, the excess and the spread distributions associated with $V(\cdot)$ (see [6]).

Write $\rho\,(=\lambda EV)$ for the data arrival rate to the system. We take the capacity of the link to be unity hence ρ is also the normalized "load" on the link. We assume that $\rho < 1$.

We use the notation $f(t) \sim_{t-t_0} g(t)$ to mean $\lim_{t-t_0}(f(t)/g(t)) = 1$.

II. Analysis of the Stationary Link Buffer Process

In this section an explicit expression for the contribution of an ongoing file to the link buffer occupancy is obtained in terms of the file size distribution and the quantities associated with the AWP [see (2)]. This is combined with the number of ongoing transfers to obtain the distribution of the buffer occupancy.

A. Observations Used in the Analysis

i) The amount of data in the link buffer at any time t is the sum of the windows from all the ongoing file transfers.
ii) Owing to the assumption that there is no loss, an AWP follows a known window increase schedule. This enables us to determine the window (which is also the session's contribution to the link buffer occupancy) for a given age.

Owing to the PS service at the file level, the distribution of the stationary number of ongoing transfers, N, is given by $P\{N = n\} = (1-\rho)\rho^n$ irrespective of the file size distribution and the AWP used. Conditioned on the number of ongoing transfers at t, the ages of the various ongoing transfers are independent; further, the age of an ongoing transfer is uniformly distributed in the interval $[0, v]$ where v is the total size of the ongoing transfer (which, in the stationary system, has distribution $V_s(\cdot)$). See [7] for these results.

Considering the stationary system, denote by G_i the window of the ith ongoing transfer. As G_i is a function only of the ith ongoing transfer's total file size and age, conditioned on N, the G_i's are i.i.d. random variables. Hence in the stationary system, the total data in the link buffer is

$$Q = \sum_{i=1}^{N} G_i. \tag{1}$$

Thus the stationary link buffer occupancy is actually a random sum of i.i.d. random variables with common distribution, say, $G(\cdot)$. As N has geometric distribution, Proposition 2.9 of [8] tells us that $Q(\cdot)$ corresponds to a sub-exponential distribution (see [8]) iff $G(\cdot)$ does so. In particular, $Q^c(\cdot)$ is regularly varying with parameter β iff so is $G^c(\cdot)$.

For an AWP, let $w(u)$ be the window of an infinitely long file with age u (that has suffered no packet loss). Assume that $w(u)$ is nondecreasing in u; Fig. 3 depicts one such instance. The contribution to the link buffer of a file of size v and age u is $G = \min(w(u), v-u)$; this is because when the age of file is u, the unacknowledged data of the file can be at most $v-u$. Define $\psi(x) := \inf\{u : w(u) > x\}$. Note from Fig. 3 that $\psi(x)$ is the age at which a window of x is achieved by an infinitely long file. It can be easily seen from the figure that, if $w(u)$ is nondecreasing, the window (and hence contribution to the link buffer) of a file of size v is $\geq x$ iff $v \geq \psi(x) + x$ and the age of the file is $u \in (\psi(x), v - x)$.

Thus, the probability that the window of an ongoing transfer is at least x is the probability that, conditioned on its file size being v, the age of the transfer is in interval $(\psi(x), v-x)$; i.e.,

$$G^c(x) = \int_{v=\psi(x)+x}^{\infty} \int_{u=\psi(x)}^{v-x} \frac{du}{v} dV_s(v) = V_e^c(x + \psi(x)). \tag{2}$$

We know that V_e is light-tailed iff V is light-tailed and also that if V is heavy-tailed then the tail of V_e is heavier than that of V (see [8, Sect. III],). Now, as $x \leq x + \psi(x)$, it follows that $G^c(x) = V_e^c(x + \psi(x)) \leq V_e^c(x)$; thus the tail behavior of $G^c(\cdot)$ is no worse than that of $V_e^c(\cdot)$.

$G(\cdot)$ for TCP-SS Controlled Transfer of Pareto Files[1]: A unit amount of data acknowledged results in a window increase by unity irrespective of the current window size. Thus $(d/du)w(u) = 1$ with $w(0) = 1$, i.e., $w(u) = u + 1$, and $\psi(x) = \max(0, x-1)$. Hence

$$G^c(x) = V_e^c(x + \psi(x)) \sim_{x\to\infty} \frac{1}{\alpha 2^{\alpha-1} x^{\alpha-1}}.$$

Note that the tail behavior of $G^c(\cdot)$ is same as that of $V_e^c(\cdot)$ which, as already observed, is the worst possible behavior of $G^c(\cdot)$. Thus the worst possible tail behavior is indeed achieved by the aggressive nature of TCPs slow start. Note that $G^c(\cdot)$, and hence $Q^c(\cdot)$, are regularly varying with parameter $(\alpha - 1)$.

[1] $V^c(v) = \min(1, (1/v^\alpha))$ and $V_e^c(v) = \min(1, (1/\alpha v^{\alpha-1}))$

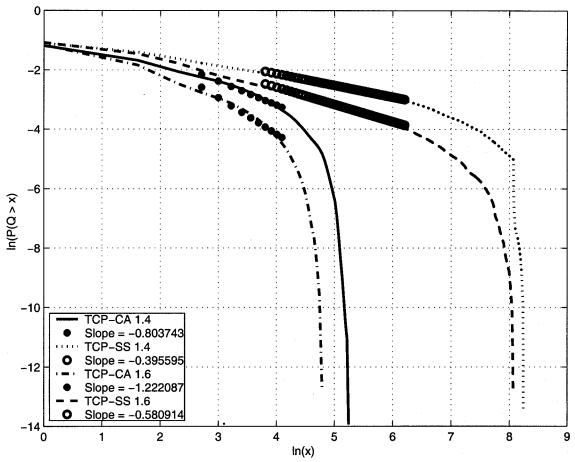

Fig. 4. Plot showing $\ln Q^c(x)$ vs $\ln(x)$ obtained from ns simulations for Pareto files with $\alpha = 1.6$ and $\alpha = 1.4$ controlled using TCP-CA and TCP-SS. The link load was set to 0.4. The link buffer uses DRR packet scheduling.

$G(\cdot)$ for TCP-CA Controlled Transfer of Pareto Files: When the window size is w, a unit amount of data acknowledged results in a window increase by $1/w$. Thus $(d/du)w(u) = (1/w(u))$ with $w(0) = 1$, i.e., $w(u) = \sqrt{2u+1}$ and thus $\psi(x) = \max(0, (x^2 - 1/2))$. Hence

$$G^c(x) = V_e^c(x + \psi(x)) \sim_{x\to\infty} \frac{2^{\alpha-1}}{\alpha x^{2\alpha-2}}.$$

Thus $G^c(\cdot)$, and hence $Q^c(\cdot)$, are regularly varying with parameter $2(\alpha-1)$. Note that, for $\alpha > 1$, it follows that $2(\alpha-1) > \alpha-1$ and hence the tail of the link buffer occupancy distribution is lighter than its worst possible behavior (which is, as seen for TCP-SS, achievable).

An AWP for Which $G(\cdot)$ has Exponential Tail for Pareto Files: Consider $w(u) = 1 + \ln(u+1)$ i.e., $\psi(x) = \max(0, e^{x-1} - 1)$. Then

$$G^c(x) = V_e^c(x + \psi(x)) \sim_{x\to\infty} \frac{e^{-(\alpha-1)x}}{\alpha e^{1-\alpha}}.$$

Thus $G^c(\cdot)$, and hence $Q^c(\cdot)$, have exponentially decaying tails.

Remark: In all three different AWPs discussed above, it can be shown that the process corresponding to the instantaneous traffic arrival rate into the link buffer at time t (denoted $a(t)$; see Figs. 1 and 2) is LRD for Pareto distributed file transfer volumes and for file transfer request arrival rates $\lambda < \lambda^*$, for some $\lambda^* > 0$ (See [9]). Even though the traffic into the link buffer is LRD, the very different buffer behavior depending on the AWP controlling the transfers emphasizes the importance of a closed-loop analysis.

B. Simulation Experiments

In Fig. 4 we plot $\ln Q^c(x)$ versus $\ln(x)$ obtained from *ns* simulations for the transfer of Pareto distributed files with $\alpha = 1.4$ and 1.6 using the TCP-CA and TCP-SS protocols; the normalized offered load ρ was set to 0.4 and the link buffer implements DRR scheduling. The plot also shows the linear approximations to curves for large values of x where the curve is roughly linear. The sharp drops observed at the ends of the curves are due to the finite simulation run lengths and are not considered in the linear approximation. The slopes of these approximations (shown in the legend) are seen to be close to their respective values predicted by the above analysis; for example, the slope for TCP-CA with $\alpha = 1.4$ is -0.8037, close to $2(\alpha-1)$ obtained from the analysis. The plot also confirms the results of Section II that with TCP-CA the tail of link occupancy distribution is lighter than that for TCP-SS. Note that the tail behavior for TCP-SS controlled transfer of Pareto 1.6 files is worse than that for TCP-CA controlled transfer of Pareto 1.4 files. Variance-time plots obtained for the $a(t)$ process shows that the $a(t)$ process is LRD with Hurst parameter $(3 - \alpha/2)$ irrespective of TCP-SS or TCP-CA as proved in [9].

III. Conclusion

We analyzed the stationary behavior of the bottleneck link buffer occupancy under AWP-controlled transfer of randomly arriving finite volume files. The fixed propagation delay across the link is negligible. The most important example of an AWP is TCP. An explicit expression for the tail of the stationary distribution of the link buffer occupancy was obtained, and was seen to have a dependence on the AWP and the file size distribution.

It was shown that with an AWP that grows its window in the way that TCP does in its congestion avoidance phase, the tail of the link buffer occupancy distribution is lighter as compared to that with an AWP that follows the window growth of the slow start phase of TCP. This is important in the view of results of [9] which proves that the traffic process into the link buffer is LRD for both the above AWP's and that an open loop analysis of a queue fed with LRD traffic as in [3] would not reflect this lightening effect. Note that most of the data of large files in the Internet are sent using the congestion avoidance phase of TCP which, as shown here, results in better behavior of link buffers even in the absence of packet losses. The presence of packet losses should result in further lightening of the tail behavior of buffer occupancy distribution. We conclude that an open-loop analysis leads to pessimistic conclusions about the buffer behavior in the Internet.

The characterization of the tail of link buffer occupancy we have developed could lead to an explanation of the sensitivity with distribution of TCP throughput performance with finite volume transfers (as observed in [10]).

The work reported in this paper is for a zero propagation delay link. In ongoing work we are studying how the results presented in this paper change as the propagation delay increases, and if random packet loss is introduced.

References

[1] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic (Extended Version)," *IEEE/ACM Trans. Networking*, Feb. 1994.

[2] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Trans. Networking*, vol. 5, no. 6, pp. 835–846, Dec. 1997.

[3] B. Tsybakov and N. D. Georganas, "Self-similar traffic and upper bounds to buffer-overflow probability in an ATM queue," *Perform. Eval.*, 1998.

[4] S. Floyd and V. Paxson, "Difficulties in simulating the internet," *IEEE/ACM Trans. Networking*, vol. 9, pp. 392–403, 2001.

[5] M. Shreedhar and O. Varghese, "Efficient fair queuing using deficit round robin," *IEEE/ACM Trans. Networking*, 1996.

[6] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[7] F. P. Kelly, *Reversibility and Stochastic Networks*, New York: Wiley, 1979.

[8] K. Sigman, "A primer on heavy-tailed distributions," *Queueing Systems, Theory and Applications*, vol. 33, 1999.

[9] A. A. Kherani and A. Kumar. Closed loop analysis of the bottleneck buffer under adaptive window controlled transfer of HTTP-like traffic. presented at Proc. IEEE Infocom. [Online]. Available: http://ece.iisc.ernet.inr~anurag/pubm.htnsl

[10] ——, "Stochastic models for throughput analysis of randomly arriving elastic flows in the internet," in IEEE Infocom, New York, June 2002.