Buffer Sizing for Nonhomogeneous TCP Sources

Lachlan L. H. Andrew, *Member, IEEE*, Tony Cui, Jinsheng Sun, Moshe Zukerman, *Senior Member, IEEE*, King-Tim Ko, *Member, IEEE*, and Sammy Chan, *Member, IEEE*

***Abstract*— Considering a single bottleneck model of TCP with a constant number of greedy and nonhomogeneous sources, and assuming that TCP timeouts do not occur, we establish necessary and sufficient conditions, related to the bandwidth delay product, to guarantee that the buffer will never empty. We also demonstrate by simulation that weaker conditions could be adequate to maintain high utilization in practice, and discuss delay and packet drop rate implications.**

***Index Terms*— Congestion control, buffer dimensioning.**

I. Introduction

SIZING of buffers in routers is an important problem. Large buffers may lead to unacceptable packet delay, while small buffers may cause excessive packet loss and inefficiency. Inefficiency due to small buffers is especially relevant if the traffic is supported by the transmission control protocol (TCP). TCP Reno [7], [10] reacts to congestion (packet loss) by halving its congestion window ($cwnd$). If buffers are too small, this often empties the queue and causes the link not to be *maximally utilized*, i.e., there are packets to be sent but the link is idle. To overcome this problem, Internet routers nowadays are designed with large buffers. To be specific, these buffers are designed to be larger than the bandwidth-delay product [6], [12]. This paper considers a simple model of TCP with drop-tail buffer management, and studies tradeoffs between buffer size (delay) and throughput.

We focus on TCP with drop-tail because: 1) most data traffic nowadays is TCP based, and 2) despite many proposals for active queue management (AQM) schemes (see [3], [8], [9], [15], [17] and references therein), drop-tail is still very popular.

The TCP protocol has been extensively studied (see, for example, [1], [4], [13]). It is well known that, for a single flow, a necessary and sufficient condition for the so-called "queue never empties" condition [4] is that the bottleneck buffer be larger than the bandwidth delay product. In this paper, we consider a single bottleneck carrying multiple flows and show in Section II that the corresponding necessary and sufficient condition, for long-lived flows, is that the buffer at the bottleneck should be larger than the *maximal* bandwidth delay product (rather than, say, the mean bandwidth delay product). However, in practice, a strict guarantee that the buffer will not empty may not be necessary, and a smaller buffer is desirable to guarantee lower delay. Consequently, we investigate by simulation in Section III the relationship between buffer size and link utilization, in the presence of transient flows.

Manuscript received October 28, 2004. The associate editor coordinating the review of this letter and approving it for publication was Dr. Nikos Nikolaou. This work was partially supported by a grant from the Research Grants Council of the Hong Kong SAR, China (Project Number CityU 1031/01E), a grant from City University of Hong Kong (Project No. 7001584) and by the Australian Research Council (ARC).

L. L. H. Andrew, T. Cui, and M. Zukerman are with the ARC Special Research Centre for Ultra-Broadband Information Networks, an affiliated program of National ICT Australia, EEE Dept., The University of Melbourne, Vic. 3010, Australia (e-mail: lha@ee.mu.oz.au).

J. Sun is with the Department of Automation, Nanjing University of Science and Technology, Nanjing, 210094, China.

K.-T. Ko and S. Chan are with the Department of Electronic Engineering, City University of Hong Kong.

Digital Object Identifier 10.1109/LCOMM.2005.06029.

II. Modelling and the Fundamental Bound

Consider a simple discrete system model with n TCP connections based on the network topology shown in Fig. 1.

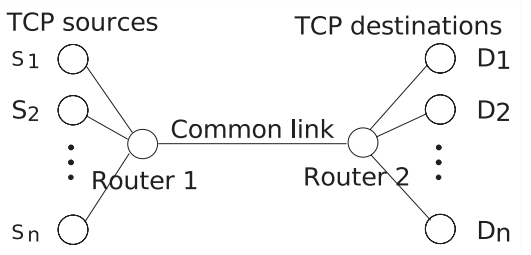

Fig. 1. The single bottleneck topology

The following assumptions are made.

1) The capacity of the link from Router 1 to Router 2, denoted μ, is the bottleneck for each connection.
2) Each flow passes through no other bottleneck links.
3) Only congestion, not channel errors, causes data loss.
4) Router 1 has a single drop-tail buffer of size B.
5) All sources are greedy (always have data to transmit).
6) When the buffer is full, each source experiences at most one packet loss.
7) As it is commonly assumed, only the congestion avoidance phase of TCP congestion control is considered.

In TCP's congestion avoidance phase, whenever the source receives $cwnd$ ACKs, the $cwnd$ increases by 1 and $cwnd+1$ packets are then sent. Upon packet loss, the TCP source decreases the $cwnd$ by half.

If it is assumed that, when a buffer overflow occurs, at most one packet is lost from any one source and no timeout occurs, then an expression can be found for the minimum buffer required to prevent the link from being starved.

Consider n sources sharing a bottleneck link, and bottlenecked nowhere else, such as in Fig. 1. This ensures that the round trip time, except queuing for the link being considered, remains constant. Let τ_i be the round trip time of the ith source minus the queuing delay. Then a tight lower bound on the queue size is

$$Q_{\min} = \left(\frac{B - \mu \max_i(\tau_i)}{2} \right)^+, \tag{1}$$

where $x^+ = x$ if $x \geq 0$, and 0 otherwise. To establish this result, consider a network in which $\tau_1 \geq \tau_j$ for all $j \neq 1$ and with a buffer size B.

To see that (1) is a bound, note that (a) aside from small statistical fluctuations, the buffer occupancy only reduces after a packet is lost, (b) the reduction is by half of the aggregate window size of flows which lose packets, and (c) the buffer occupancy is B before the packet loss. Let w_j be the window size of flow j. The reduction in buffer occupancy is maximised when a loss occurs in each flow, and so it is maximised when the aggregate of all flows' windows is maximised. This occurs when $w_j \approx 0$ for all $j \neq 1$ and $w_1 \approx \mu\tau_1 + B$. If

$$B > \mu \max_i(\tau_i), \tag{2}$$

then the residual buffer occupancy is $w_1/2 - \mu\tau_1 = Q_{\min}$. Otherwise, it is 0, which is again $Q_{\min}$.

To see that (1) is a tight bound, note that it is possible (albeit unlikely) for losses to occur in such a way that the above scenario ($w_j \approx 0$ for all $j \neq 1$) occurs.

The implication of the bound (1) is that, if (2) is satisfied, then $Q_{\min} > 0$ and the link can never be starved of packets. It thus achieves full utilisation.

We can therefore conclude that for the single bottleneck case we considered, the buffer being larger than the maximal bandwidth delay product, is a necessary and sufficient condition for the buffer never to empty.

This new result is consistent with results of other TCP analyses [1], [5], [13]. Note also that by (1), Q_{min} is independent of n, the number of TCP connections. However, we may expect that the distribution of the buffer occupancy does depend on n, with the tail becoming increasingly light as n increases. This can be justified by noticing that the maximum decrease in buffer occupancy only occurs when all n flows simultaneously halve their windows. As n increases, the probability of this decreases.

III. Smaller buffers and transient flows

Although the above condition is necessary to guarantee full link utilisation, the chance of the buffer emptying with a smaller buffer may be very low. Let $\bar{\tau}$ be the mean round trip time. In the case of homogeneous sources, $\tau_i = \bar{\tau}$ for all i, the maximum and mean values of τ_i coincide. Namely, $\max_i(\tau_i) = \bar{\tau}$. Thus in this case, according to (2), dimensioning the buffer to be equal to the *mean bandwidth delay product*, defined by $\mu\bar{\tau}$, is sufficient. However, in this case, the flows can become synchronized causing windows to halve simultaneously. When round trip times are nonhomogeneous, it is expected that only a small subset of flows will halve their windows at any one time, leading to reduced buffering requirements. This suggests that, also in the general case of nonhomogeneous sources, it may be sufficient to use

$$B > \mu\bar{\tau}. \tag{3}$$

The simulation results presented in [16] demonstrate that this is indeed sufficient for the cases examined.

We will now demonstrate by simulations that even buffers smaller than the average bandwidth delay product can lead to very small proportion of time where the buffer is empty.

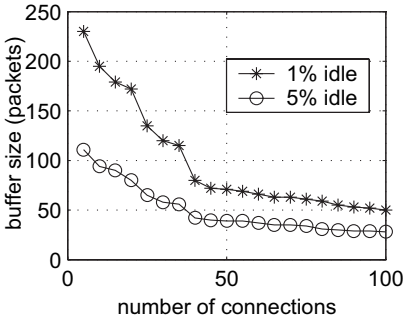

Fig. 2. Minimal buffer required to achieve a given proportion of idle time as a function of the number of sources.

We use the topology of Fig. 1 and the parameters: $\mu = 4,000$ packet/s (i.e., about 16 Mbit/s for packet size of 500 bytes). The mean propagation delay ($\overline{\tau}$) is 100 ms, giving a bandwidth delay product of 400 packets. The number of sources (n) and the buffer size (B) are varied. The actual propagation delays are governed by a Pareto distribution with mean 100 ms. To this end, we considered the propagation delay to be a Pareto random variable Ψ. In particular, we generated n deviates from a Pareto distribution with parameters γ and δ defined by its complementary distribution function given by

$$P(\Psi > x) = \begin{cases} \left(\frac{x}{\delta}\right)^{-\gamma}, & x \geq \delta \\ 1, & \text{otherwise.} \end{cases}$$

Its mean is given by $E[\Psi] = \delta\gamma/(\gamma - 1)$.

In our simulation experiments, we chose $\gamma = 1.2$, which implies that $\text{Var}[\Psi] = \infty$. To fit the mean $E[\Psi] = \bar{\tau} = 100$ ms requires $\delta = 50/3$. This provides a good match to measurements [11]. Each simulation run lasted for 1000 s.

An interesting question is what is the minimal buffer size for the buffer to be empty for an acceptable proportion of time. Fig. 2 presents *ns-2* simulation [14] results that answer this question for the case of long-lived flows. The number of sources n is varied between 1 and 100. For each n value, we run many simulations, each for a different value of the buffer size (B) to find the minimal value of B such that the proportion of time the buffer is empty is first 1% and then 5%.

This graph shows that buffers very much less than the bandwidth delay product (of 400 packets) are sufficient to yield high throughput when the number of flows is significant. Even for $n = 5$ users, a throughput of 99% is obtained using a buffer size of slightly over half the bandwidth delay product. This is consistent with recent analytic results [2].

To investigate the impact of finite flow duration, we simulated a system with an Engset (finite population) arrival process, and sessions of length uniformly distributed from 1 to 200 packets (with mean 100.5). After each transfer, the source waited an exponential time, with mean 2 seconds, before starting a new session. Fig. 3 shows the percentage of idle time as a function of the buffer size for populations of 200 and 500 users. This yields on average around 110 (200 users) and 400 (500 users) simultaneous flows, although the actual average is a decreasing function of the buffer size.

Smaller buffers also mean a higher packet drop rate. In Fig. 4, we present the packet drop rate versus the buffer size

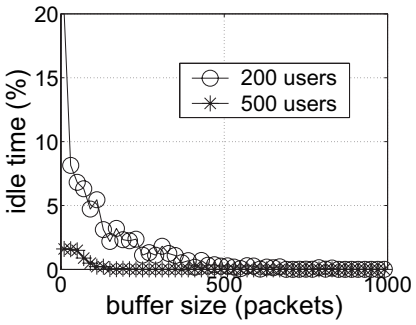

Fig. 3. Link idle time versus the buffer size for populations of 200 and 500 users (on average around 110 and 400 active sources).

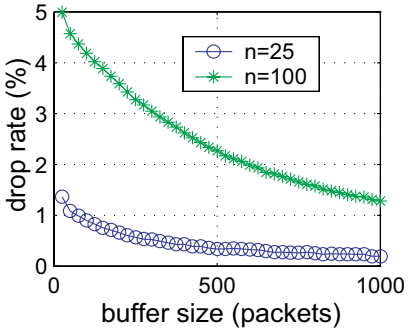

Fig. 4. Packet loss probability versus the buffer size for $n = 25, 100$ sources.

for $n = 25$ and $n = 100$ long-lived flows. As expected, the drop rate decreases as the buffer size increases. However, in contrast to the throughput, the drop rate is significantly worse when a larger number of sources share the buffer. To see this, note that the average window size is smaller when there are more flows, and so the reduction in packet arrival rate caused by one flow halving its window is less. Thus, more halving events need to occur, requiring a higher packet loss rate.

Even though the foregoing results show that buffers significantly smaller than the bandwidth delay product may be sufficient, it could be argued that memory is cheap, and so even a slight performance improvement would justify using large buffers. However, the primary benefit of using smaller buffers is a reduction in the queuing delay. This is quantified in Fig. 5, which shows *ns-2* simulation results for the mean percentage of buffer occupancy versus the buffer size for $n = 25$, $n = 100$ and $n = 1000$ long-lived flows.

The mean fractional occupancy is significantly larger when the number of users is larger. That is again because the window size of each flow is smaller, and so the amount by which the occupancy decreases on each packet loss is smaller. It is this increase of mean occupancy with number of connections which explains the results in the two previous figures.

IV. Conclusions

We have established that the buffer size being at least the bandwidth delay product is necessary and sufficient condition to guarantee that the buffer never empties for a single bottleneck model of TCP with a constant number of saturated and nonhomogeneous sources and assuming that TCP timeouts do not occur. For this model. we have demonstrated by simulation that smaller buffers are adequate to maintain high utilization in

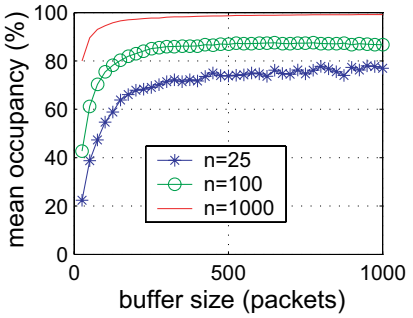

Fig. 5. Mean percentage of buffer occupancy versus the buffer size for $n = 25$, $n = 100$ and $n = 1000$ sources.

practice. We have also demonstrated by simulation the increase of delay and drop rate with an increase in the number of sources, and that delay increases and drop rate decreases with an increase in buffer size.

References

[1] E. Altman, F. Boccara, J. Bolot , P. Nain , P. Brown, D. Collange, and C. Fenzy, "Analysis of the TCP/IP flow control mechanism in high-speed wide-area networks," in *Proc. 34th IEEE Conference on Decision and Control*, Dec. 1995, pp. 368–373.

[2] G. Appenzeller, I. Keslassy, and N. McKeown, "Sizing router buffers," in *Proc. ACM SIGCOMM*, 2004, pp. 281-292.

[3] S. Athuraliya, V. H. Li, S. H. Low, and Q. Yin, "REM: active queue management," *IEEE Network*, vol. 15, pp. 48–53, May/June 2001.

[4] P. Brown, "Resource sharing of TCP connections with different round trip times", *Proc. IEEE INFOCOM 2000*, Mar. 2000, pp. 1734–1741.

[5] S. G. Choi, R. Mukhtar, J. K. Choi, and M. Zukerman, "Efficient macro mobility management for GPRS IP networks," *J. Commun. and Networks*, vol. 5, pp. 55–64, Mar. 2003.

[6] Cisco 12000 Series Gigabit Switch Router (GSR) Gigabit Ethernet Line Card (online). Available: http://www.cisco.com/warp/public/cc/pd/rt/12000/prodlit/

[7] K. Fall and S. Floyd, "Simulation-based comparisons of Tahoe, Reno and SACK TCP", *ACM Comp. Commun. Review*, vol. 26, pp. 5–21, July 1996.

[8] W. Feng, D. Kandlur, D. Saha, and K. Shin, "The Blue Active Queue Management Algorithms", *IEEE/ACM Trans. Netw.*, vol. 10, pp. 513–528, Aug. 2002.

[9] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Trans. Networking*, vol. 1, pp. 397–413, Aug. 1993.

[10] V. Jacobson, "Congestion avoidance and control," *Proc. ACM SIGCOMM '88*, 1988, pp. 314–329.

[11] H. Jiang and C. Dovrolis, "Passive estimation of TCP round-trip times," *ACM Comp. Commun. Review*, vol. 32, pp. 75–88, July 2002.

[12] Juniper M-series Routers, http://www.juniper.net/products/dsheet/100042.html#01

[13] T. V. Lakshman and U. Madhow, "The performance of TCP/IP for networks with high bandwidth-delay products and random losses", *IEEE/ACM Trans. Networking*, vol. 5, pp. 336–350, June 1997.

[14] The network simulator ns-2 (online). Available: http://www.isi.edu/nsnam/ns/

[15] J. Sun, K.-T. Ko, G. Chen, S. Chan, and M. Zukerman, "PD-RED: to improve the performance of RED," *IEEE Commun. Lett.*, vol. 7, pp. 406–408, Aug. 2003.

[16] J. Sun, M. Zukerman, K.-T. Ko, G. Chen, and S. Chan, "Effect of large buffers on TCP queueing behavior," in *Proc. IEEE INFOCOM 2004*, Mar. 2004, pp. 751–761.

[17] B. Wydrowski and M. Zukerman, "QoS in best-effort networks," *IEEE Commun. Mag.*, vol. 40, pp. 44–49, Dec. 2002.