A statistical bandwidth sharing perspective on buffer sizing

J. Augé, and J. Roberts

France Telecom
DRD/CORE/CPN
38, rue du Général Leclerc
92794 Issy-Moulineaux, France
{jordan.auge, james.roberts}@orange-ftgroup.com

Abstract. The issue of buffer sizing is rightly receiving increasing attention with the realization that the bandwidth delay product rule-of-thumb is becoming unsustainable as link capacity continues to grow. In the present paper we examine this issue from the light of our understanding of traffic characteristics and the performance of statistical bandwidth sharing. We demonstrate through simple analytical models coupled with the results of ns2 simulations that, while a buffer equivalent to the bandwidth delay product is certainly unnecessary, the recently advocated reduction to a few dozen packets is too drastic. The required buffer size depends significantly on the peak exogenous rate of multiplexed flows.

1 Introduction

The rule-of-thumb whereby router buffers should be sized to store a full bandwidth delay product of data has recently come under considerable scrutiny [1,2,3,4,5,6,7]. It has been pointed out that to realize such large buffers for future 40Gb/s links is a significant design challenge for electronic routers and remains completely impractical for future optical routers. Moreover, the original reasoning behind the rule-of-thumb [8] is no longer valid for the present Internet backbone, both with respect to the size of links and their traffic and to the relative costs of memory and bandwidth.

Realized performance, in terms of packet loss and delay and flow throughput, for a given buffer size clearly depends on the assumed characteristics of link traffic. The previously cited papers differ significantly in their assumptions and consequently arrive at some conflicting conclusions. Our aim in the present paper is to identify the essential components of a typical mix of flows and to evaluate the buffer size performance trade-off under realistic traffic assumptions.

Internet traffic is composed of a dynamic superposition of finite size flows. An important characteristic of the flows sharing a given link is their exogenous "peak rate". This is the rate the flow would achieve if the link were of unlimited capacity. Some flows with a high peak rate will be bottlenecked by the considered link and will share bandwidth using end-to-end congestion control. However, the vast majority of flows are not bottlenecked because their peak rate, determined for instance by a low speed access line, is much smaller than the current fair share. The number of bottlenecked flows is not an exogenous traffic characteristic but results from the dynamic statistical bandwidth sharing process and can be characterized as a function of the overall link load.

In the paper we review some simple models of statistical bandwidth sharing and identify two main link operating regimes relevant for buffer sizing. These are a "transparent" regime, where no flows are bottlenecked, and an "elastic" regime where a relatively small number of bottlenecked flows share bandwidth with a background traffic produced by many non-bottlenecked flows. While small buffers are sufficient in the transparent regime, it appears necessary to scale buffer size with link capacity in the elastic regime. We begin by reviewing the existing literature on the buffer sizing issue.

2 Related work on buffer sizing

Appenzeller and co-authors were first to argue that the bandwidth delay product rule-of-thumb was both unsustainable, given the anticipated increase in network link capacity, and unnecessary [1]. Assuming flow congestion window (`cwnd`) evolutions are desynchronized they argued the required buffer should be proportional to $\sqrt{N}$ where N is the number of flows. It was noted, however, by Raina and Wischik [2] and Raina et al. [4] that the occurrence of flow synchronization depends on the buffer size and that, for a very large number of flows, the buffer size proposed in [1] is still too big. They suggest the buffer capacity needs to be as small as a few tens of packets. Instability was not observed in [1] because the authors only performed simulations for a few hundred flows whereas the phenomenon occurs for some thousands. In Section 3 we argue that it is not reasonable to suppose so many flows are actually *bottlenecked* at the link and therefore question the validity of this argument in favour of very small buffers.

Dhamdhere et al. [6] recognize the importance of distinguishing bottlenecked flows and non-bottlenecked flows. They suggest it is necessary to achieve a low packet loss rate while realizing high link utilization. Consequently they advocate a relatively large buffer that is proportional to the number of bottlenecked flows. The authors advance their analysis in [7], notably introducing open and closed loop dynamic flow level traffic models that correspond to our notion of statistical bandwidth sharing. However, the model in [6] is claimed to be valid when the bottlenecked flows constitute more than 80% of link load and performance is evaluated in [7] for a very high link load when some 200 bottlenecked flows are in progress. We again question the relevance of these traffic assumptions in evaluating buffer requirements.

The model proposed by Enachescu et al. [5,9] evaluates buffer requirements when none of the flows is bottlenecked. This is a valid assumption for many network links and corresponds to what we term the transparent regime (see Section 3). It is suggested in [9] that buffer size should be proportional to the log of the maximum TCP congestion window size. A necessary assumption is that packets are paced at the average rate determined by the window size rather than emitted as bursts. We note that the analysis in [2,4] is based on a fluid model and also therefore makes an implicit assumption that packet arrivals are not bursty. We believe buffer requirements must be evaluated for a mix of non-bottlenecked flows (where packets are "paced" to the flow peak rate) and bottlenecked flows that typically emit packets in bursts.

The present paper builds on our preliminary work [10]. We seek to evaluate buffer size based on our understanding of the statistical nature of traffic and accounting for the burstiness of TCP packet emissions.

3 Statistical bandwidth sharing

Consider a link of capacity C shared by a set of flows. The size and make up of this set of flows varies in time as flows of various types and characteristics arrive and depart. To understand what constitutes a typical traffic mix (e.g., for evaluating required buffer size), it is necessary to evaluate the performance of appropriate statistical bandwidth sharing models.

3.1 Processor sharing models

The processor sharing (PS) model for statistical bandwidth sharing provides insight into the way TCP flow-level performance depends on traffic characteristics [11,12]. In the PS model, flows are assumed to arrive according to a Poisson session model (a large population of users independently generate sessions, each session consisting of a succession of flows and think times) and to share link bandwidth perfectly fairly with any other concurrent flow. It can then be demonstrated that performance measures like expected flow throughput are largely insensitive to detailed traffic characteristics like the distribution of flow size or the number of flows in a user session. The essential characteristics are the mean load ρ, equal to flow arrival rate × mean flow size / link rate, and the flow "peak rate", the maximum rate a flow can attain independently of the considered link. Insensitivity is only approximate when flows have different peak rates or share the link unfairly due to different round trip times but the broad characteristics deduced from ideal symmetric models remain valid.

If flows can all individually attain the link rate, the number of flows in progress in the ideal fluid PS model has a geometric distribution of mean $\rho/(1-\rho)$. Despite the simplicity of the model, this is a good indication that the number of flows in contention at any instant would typically be quite small (e.g., less than 20 with probability .99 at 80% load). Of course, we know that the number of flows in progress on most network links is, on the contrary, very large (tens of thousands on a Gb/s link, say). This is because the vast majority of flows are peak rate limited and cannot realize a fair bandwidth share.

3.2 Throughput performance

To illustrate statistical bandwidth sharing performance, we consider the following measure of flow throughput: γ = mean flow size / mean flow duration. For fair sharing, the parameter γ can also be interpreted as the expected instantaneous throughput of a flow in progress at an arbitrary instant [13]. We consider generalized PS models where flows fairly share an overall service rate that depends on the number in progress. This service rate depends, for example, on the link buffer size. Let the service rate when i flows are in progress be $\phi(i)C$ and write $\Phi(i) = \prod_1^i \phi(i)$. We have,

$$\gamma = \frac{\sum \rho^i / \Phi(i)}{\sum i \rho^{i-1} / \Phi(i)}. \tag{1}$$

3.3 Bandwidth sharing regimes

Figure 1 plots γ as a function of ρ when flows have no peak rate constraint ($\phi(i) = 1$) and when flows all have the same peak rate $p = 0.1C$ ($\phi(i) = \min(ip, 1)$). These simple cases illustrate two important points: i) the number of bottlenecked flows is very large only when link load is close to 1 (this number is proportional to $1/\gamma$), ii) when flows are peak rate limited, the link is transparent to throughput performance up to high loads (close to $(1 - p/C)$). More generally, in discussing buffer sizing, it is useful to distinguish three bandwidth sharing regimes:

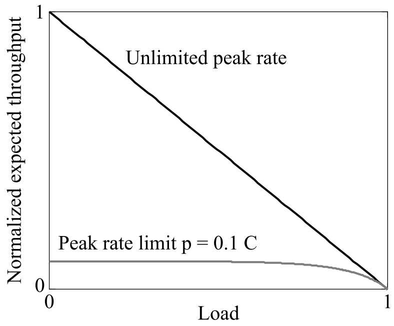

Fig. 1. Per flow throughput γ for PS model as fraction of link capacity C and as a function of load ρ

- an overload regime ($\rho > 1$) where realized flow throughput tends to zero as the number of competing flows increases,
- a transparent regime where the sum of peak rates of all flows remains less than link capacity (with high probability),
- an intermediate "elastic" regime where the majority of flows are peak rate limited but share the link with a small number of other flows capable of using all the residual capacity.

Buffer sizing is clearly inadequate for dealing with congestion in the overload regime. Alternative mechanisms to deal with this (i.e., traffic engineering and admission control) should also avoid situations of near overload when the number of bottlenecked flows grows rapidly. In the next sections we discuss buffer requirements for transparent and elastic regimes, respectively.

4 Buffer sizing for the transparent regime

The transparent regime is characterized by the fact that the sum of flow peak rates is, with high probability, less than link capacity. The buffer must be sized to avoid significant loss due to coincident packet arrivals from independent flows. We assume the rate of flows is well defined at the time scale of packet emissions as, for instance, when it is limited by an upstream access line.

4.1 Locally Poisson arrivals

Figure 2 depicts the overall rate of a superposition of peak rate limited TCP Reno flows using the simulation set-up specified in Figure 3. Without bottlenecked traffic, the flows arrive as a Poisson process and have a size drawn from an exponential distribution of 100 packet mean. The figures plot the average rate in successive 100ms slots. The rate is shown for two flow peak rates, $p = 200$Kb/s and $p = 50$Kb/s, and that measured for a Poisson process.

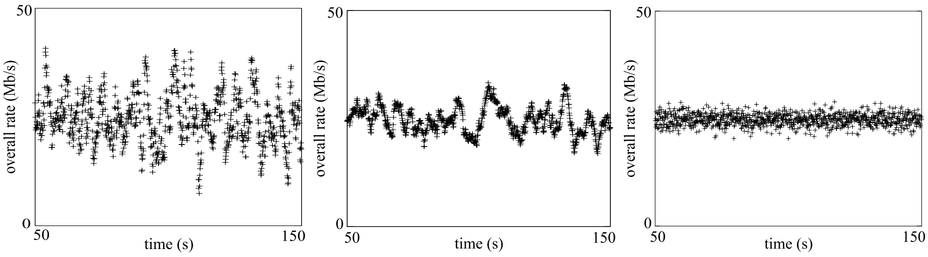

Fig. 2. Packet arrival process: rate of packets arriving in successive 100ms slots for flows of peak rate $p = 200$Kb/s, $p = 50$Kb/s and $p = 0$.

Visibly, a Poisson packet arrival process is not a good approximation unless the peak rate is very small relative to the link rate. However, in a small time interval (e.g., in each 100ms slot), the packet arrival process, as a superposition of a large number of periodic processes, is approximately Poisson. The figure depicts a realization of this modulated Poisson process. We denote its intensity by Λ_t.

Assuming Poisson arrivals allows buffer occupancy to be approximated locally by that of the M/G/1 queue. To simplify, one can further assume exponential packet sizes and approximate packet loss probability for a buffer of size B by $(\Lambda_t/C)^B$.

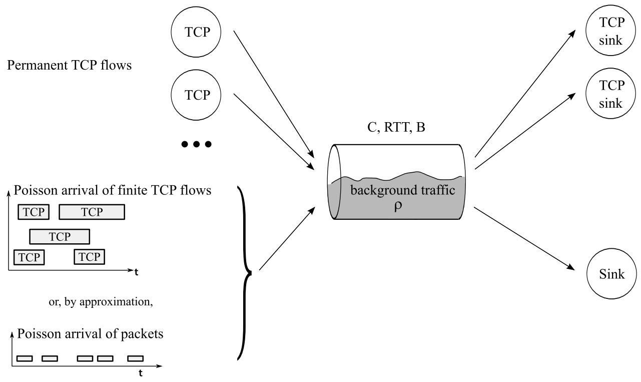

Fig. 3. Simulation set-up: unless otherwise stated, simulations use the following base set of parameters: $C = 50$ Mb/s, $B = 20$ packets, $RTT = 100$ ms, $\rho_b = 0.5C$, FIFO scheduling, 1000 bytes packets.

4.2 Required buffer size

A possible approach for buffer sizing is to compute an average loss rate by conditioning on the distribution $F(\lambda)$ of Λ_t and requiring $\int(\lambda/C)^B dF(\lambda) < \epsilon$. This is reasonable when the rate variations are rapid so that ϵ is a good measure of the performance of any given flow. It turns out that, for a peak rate less than $.1C$ and $\epsilon > .001$, the required buffer size is the same as would be required for the Poisson packet arrival process. In other words, the M/M/1 formula $\rho^B < \epsilon$ is a useful sizing guideline. For example, a buffer of 20 packets limits admissible load to $\rho = .79$ for $\epsilon = .01$ or $\rho = .7$ for $\epsilon = .001$.

5 Buffer sizing for the elastic regime

In general, there is no mean to guarantee flow peak rates are limited and it is therefore important to understand the impact of buffer size on performance in the elastic regime (i.e., when one or several bottlenecked flows combine with background load to momentarily saturate the link for periods that are long compared to the time scale of packet emissions).

5.1 Unlimited rate bottlenecked flows

To simplify analysis and discussion, we suppose a clear dichotomy between flows with unlimited peak rate and a background traffic composed by flows having a low peak rate. Furthermore, we assimilate the background traffic to a Poisson packet arrival process producing load $\rho_b C$. This simplification greatly facilitates the simulation experiments and reproduces the broad behavioural characteristics of more realistic background traffic.

To evaluate throughput performance we proceed as follows. For given link capacity, buffer size and background load, we successively simulate a number of permanent bottlenecked TCP flows. For each number i (between 1 and 100), we evaluate the overall realized throughput $\phi(i)$ expressed as a fraction of residual capacity $C(1-\rho_b)$. We then derive the expected flow throughput γ by formula (1). This corresponds to a quasi-stationary analysis allowing us to ignore phenomena like loss of throughput in slow start and momentary unfairness.

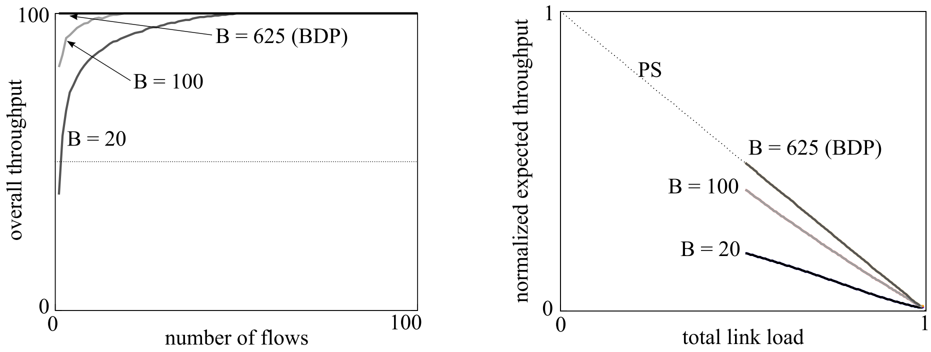

Fig. 4. Overall throughput $\phi(i)$ with i flows as fraction of residual bandwidth for a various number of flows, and expected flow throughput γ as function of load ρ, for various buffer sizes

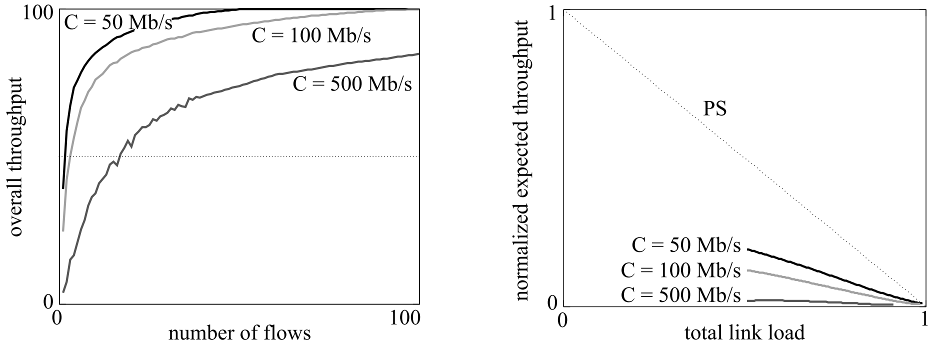

Fig. 5. Overall throughput $\phi(i)$ with i flows as fraction of residual bandwidth for a various number of flows, and expected flow throughput γ as function of load ρ, for various capacities

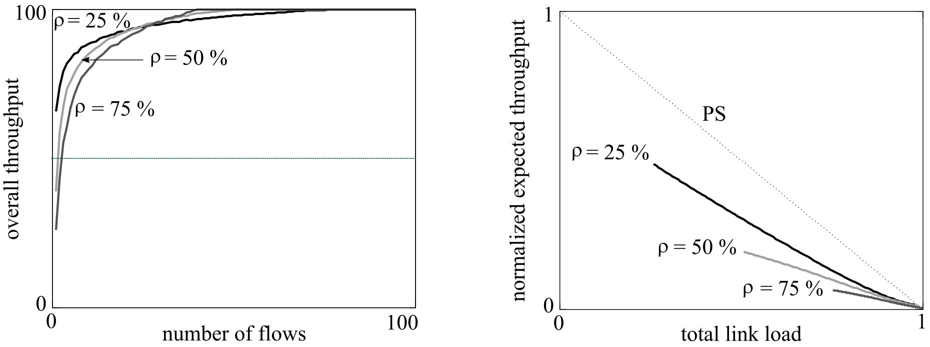

Fig. 6. Overall throughput $\phi(i)$ with i flows as fraction of residual bandwidth for a various number of flows, and expected flow throughput γ as function of load ρ, for various background loads

Figures 4, 5, 6 depict the values of $\phi(i)$ and γ as a function of link load ρ for a range of configurations. Note that γ is only defined for loads greater than the background load ρ_b and its value at that load is determined by $\phi(1)$.

The results show that there is a significant loss of throughput with small buffers (Fig. 4) and that this loss is accentuated as link capacity increases (Fig. 5). The higher the background load, the more difficult it is for the TCP flows to fully use the residual capacity (Fig. 6).

To understand the loss in throughput it is necessary to explain the behaviour with just one bottlenecked flow. This determines $\phi(1)$ and consequently the form of γ which is approximately linear, decreasing from the maximum for $\rho = \rho_b$ to 0 for $\rho = 1$.

While the TCP window is small compared to the residual bandwidth delay product $C(1-\rho_b)\times$ `RTT`, packets are emitted in bursts starting at instants separated by a Round Trip Time (`RTT`). By TCP self-clocking, the sum of the burst rate and the rate of the background traffic is very close to the link capacity. Buffer occupancy therefore tends to increase under this heavy load while the burst is in progress and then to empty when only background packets arrive. In the absence of loss, TCP increases `cwnd` by 1 packet per `RTT`, prolonging the period of overload. At some point background and bottlenecked flow packet arrivals combine to saturate the buffer and a packet is lost.

If the residual bandwidth delay product is sufficiently large and the buffer is small, the process describing the value of `cwnd` when the packet loss occurs depends only on ρ_b. This determines the average window and therefore the flow throughput in this regime. For a larger buffer size, `cwnd` is able to increase further and eventually attain the value that completely fills the residual bandwidth.

Figure 7 plots the product $\phi(1)C(1-\rho_b)$`RTT` as a function of buffer size. For small buffers this is equal to the expected value of `cwnd` and depends only on ρ_b. As the buffer size increases, the residual bandwidth is used entirely and the function flattens to a horizontal line.

The form of $\phi(1)$ as a function of B suggests the buffer should be sized to at least avoid the initial high degradation in throughput. It is not necessary, however, to attain 100% efficiency and a buffer considerably smaller than the residual bandwidth delay product would be sufficient.

A possible approach would be to set B to the value where the common curve for small buffers and given background load ρ_b intersects with the horizontal line representing the residual bandwidth delay product.

Inspection of the form of the common curve for given background load suggests an approximate dependency in B^2. In other words, the required buffer size according to the above approach would be roughly proportional to the square root of the residual bandwidth. This clearly requires further investigation, notably by more realistically modelling background traffic, but is an indication of the likely dependence of buffer size on link capacity for this elastic regime.

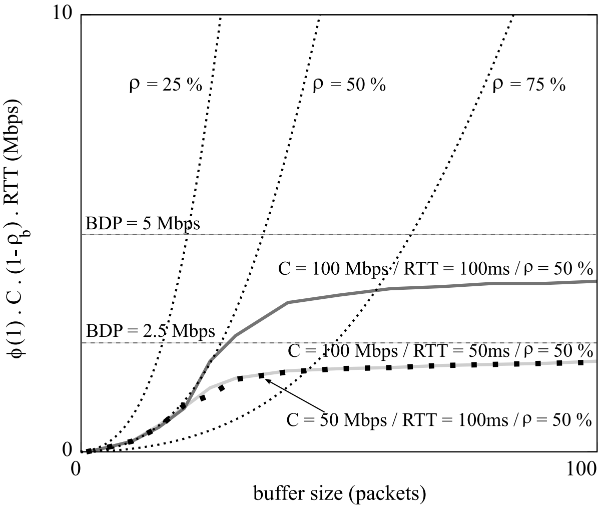

Fig. 7. Throughput performance as function of buffer size for one bottlenecked flow

5.2 Peak rate limited bottlenecked traffic

The assumption of unlimited peak rate bottlenecked flows is not necessarily reasonable since even the rate of high speed access lines is generally only a fraction of link rate. To illustrate the impact of a limited peak rate, we assume flows of peak rate p share a link with Poisson background traffic. Figure 8 plots throughput $\phi(i)$, as a function of the number of bottlenecked flows i, and γ/C, as a function of link load, for a number of configurations.

Results show that $\phi(i)$ increases linearly while the total rate of bottlenecked flows is somewhat less than the residual capacity as each flow realizes its peak rate and the link operates in the transparent regime. When overall load attains a level where the bottlenecked flows begin to lose packets, however, the inefficiency of small buffers is again apparent. For example, for $p = 2$ Mb/s in Figure 8, the throughput $\phi(i)$ dips when there are more than 11 flows and only increases to nearly 100% for a much larger number.

The efficiency loss with small buffers in this case is less significant for throughput performance, however, as illustrated by the behaviour of γ as a function of link load. The loss in throughput is only visible at high loads where it accentuates the degradation occurring in the ideal PS model (seen here with B=625).

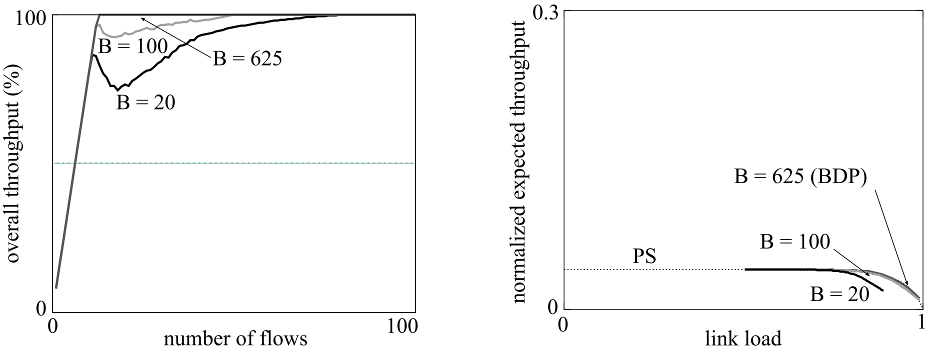

Fig. 8. Overall throughput $\phi(i)$ with i flows as fraction of residual bandwidth and expected flow throughput γ as function of load ρ, for peak rate limited bottlenecked flows (2Mb/s)

5.3 Paced TCP

In [9] it is proposed that flows that are not peak rate limited by an access line should use pacing. This would indeed attenuate the loss of throughput for small buffers arising when a window is emitted as a burst at the start of each `RTT` interval. Figures 9 illustrate the evolution of `cwnd` for a single bottlenecked flow with[1] and without pacing with $B = 20$ and $B = 100$. The results confirm that pacing significantly improves the performance of the small buffer since throughput is roughly 50% higher. The difference is negligible for the larger buffer $B = 100$.

[1] We used the paced TCP ns2 code made available by D.X. Wei : *A TCP pacing implementation for NS2*, available at `http://www.cs.caltech.edu/ weixl/technical/ns2pacing/ index.html`, with the *traditional pacing* option.

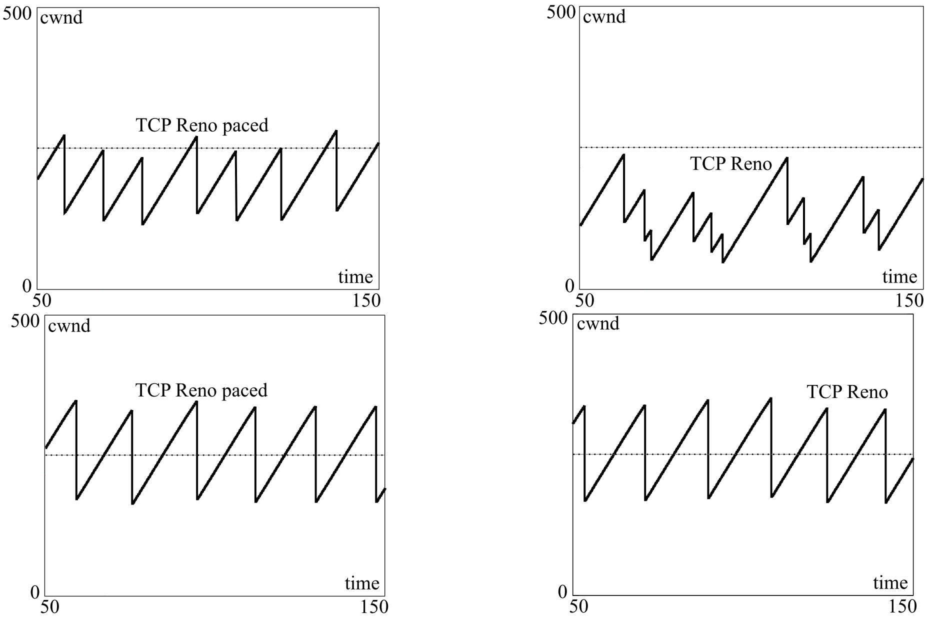

Fig. 9. Evolution of `cwnd` for one paced Reno flow (left) and one Reno flow (right), with $B = 20$ packets (top) and $B = 100$ packets (bottom), for $C = 50$ Mbps, $\rho_b = 0.5$

6 Conclusions

The relation between buffer size and realized performance clearly depends on the assumed traffic characteristics. The most significant characteristic is the mix of exogenous flow peak rates, the rates flows would attain if the considered link were of unlimited capacity. The link load (flow arrival rate × mean flow size / link capacity) then determines which, if any, high peak rate flows are bottlenecked, the remainder constituting a background load. We distinguish three main statistical bandwidth sharing regimes:

1. when all peak rates are a relatively small and load is not too close to 1, the sum of flow rates remains less than link capacity with high probability; we refer to this as the transparent regime; a simple M/M/1 queue model can be used to evaluate the relationship between buffer size and packet loss probability; a small buffer is then adequate; for example, a 20 packet buffer overflows with probability 0.01 at a load close to 80%;
2. when some flows can individually saturate the residual link bandwidth not used by the background load due to low peak rate flows, bandwidth sharing is controlled by end-to-end congestion control; we refer to this as the elastic regime; with the current practice of sending packets as soon as they are

authorized by the receipt of an acknowledgement, a small buffer tends to overflow too early to allow full development of `cwnd` and utilization can be very low; required buffer size in this regime increases with the residual bandwidth delay product; preliminary empirical evidence suggests buffer size should be proportional to the square root of the residual bandwidth delay product;

3. when the highest peak rate flows must combine (i.e., several flows in parallel) to saturate the residual bandwidth, we have a more general intermediate transparent/elastic regime; when the peak rate of the (potentially) bottlenecked flows is a relatively small fraction of the residual bandwidth (e.g., 1/10), and overall load is not too close to 1, the link is rarely saturated and a small buffer sized for the transparent regime is adequate.

Since large buffers and fair queuing appear to be impractical propositions for future optical routers, it appears important to ensure that these always operate in the transparent regime. It may be sufficient to rely on the continuing large disparity between flow peak rates and backbone link capacity, as suggested in [9].

References

1. Appenzeller, G., Keslassy, I., McKeown, N.: Sizing router buffers. Proceeding of ACM SIGCOMM '04, Portland, Oregon (September 2004)
2. Raina, G., Wischik, D.: Buffer sizes for large multiplexers: TCP queueing theory and instability analysis. NGI'05, Rome (April 2005)
3. Wischik, D., McKeown, N.: Part I: buffer sizes for core router. ACM SIGCOMM Computer Communication Review, v.35 n.3 (July 2005)
4. Raina, G., Towsley, D., Wischik, D.: Part II: control theory for buffer sizing. ACM SIGCOMM Computer Communication Review, v.35 n.3 (July 2005)
5. Enachescu, M., Ganjali, Y., Goel, A., McKeown, N., Roughgarden, T.: Part III: routers with very small buffers. ACM SIGCOMM Computer Communication Review, v.35 n.3 (July 2005)
6. A.Dhamdhere, Dovrolis, C., Jiang, H.: Buffer sizing for congested internet links. Proceedings of IEEE INFOCOM, Miami FL (March 2005)
7. A.Dhamdhere, Dovrolis, C.: Open issues in router buffer sizing. ACM SIGCOMM Computer Communications Review (editorial section) (January 2006)
8. Villamizar, C., Song, C.: High performance TCP in ANSNET. Computer Communications Review, V. 24 N. 5, October 1994, pp. 45-60 (1994)
9. Enachescu, M., Ganjali, Y., Goel, A., McKeown, N., Roughgarden, T.: Routers with very small buffers. Proceedings of the IEEE INFOCOM'06, Barcelona, Spain (April 2006)
10. Augé, J., Roberts, J.: Buffer sizing for elastic traffic. NGI'06, València (April 2006)
11. Fredj, S.B., Bonald, T., Proutière, A., Régnié, G., Roberts, J.: Statistical bandwidth sharing: a study of congestion at flow level. SIGCOMM 2001, San Diego, CA, USA (August 2001)
12. Bonald, T., Proutière, A.: Insensitivity in processor-sharing networks. Proceedings of Performance (2002)
13. Bonald, T.: Throughput performance in networks with linear capacity constraints. Proceedings of CISS 2006 (2006)