INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

TCP modeling in the presence of nonlinear window growth

Eitan Altman, Konstantin Avrachenkov, Chadi Barakat — Rudesindo Núñez-Queija

N° 4312

November 2001

THÈME 1

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--4312--FR+ENG

TCP modeling in the presence of nonlinear window growth

Eitan Altman, Konstantin Avrachenkov, Chadi Barakat * , Rudesindo Núñez-Queija †

Thème 1 — Réseaux et systèmes
Projet MISTRAL

Rapport de recherche n° 4312 — November 2001 — 19 pages

Abstract: We develop a model for TCP that accounts for both sublinearity and limitation of window increase. Sublinear window growth is observed when the round-trip time of the connection increases with the window size. The limitation is due to the window advertised by the receiver. First, we derive the required conditions for the stability of the model. Then, we write the Kolmogorov equation under Markovian assumptions. The model is solved 1analytically for some particular cases. A good match between the throughput predicted by the model and the throughput measured on real TCP connections is reported.

Key-words: TCP, non-linear window growth, Kolmogorov equations

* INRIA, Sophia Antipolis, France, Email: {altman,k.avrachenkov,cbarakat}@sophia.inria.fr
† CWI, Amsterdam, The Netherlands, Email: sindo@cwi.nl, also with the Eindhoven University of Technology.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

La modelisation de TCP dans le cas de la croissance sous-linéaire de la fenêtre

Résumé : Nous developpons un modele de TCP qui tient compte de la sous-linéarité et de la limitation de la croissance de la fenêtre. L'evolution sous-linéaire de la fenêtre peut être expliquée par la relation entre le débit instantané d'une connexion et le RTT (Round Trip Time) associé. La limitation de la fenêtre est imposée par le récepteur. Tout d'abord, nous établissons les conditions de stabilité pour le modele. Nous obtenons alors les équations de Kolmogorov qui nous permettent de résoudre analytiquement le modele dans quelques cas particuliers. Ces résultats analytiques ont été ensuite validés par des mesures sur des connexions réelles.

Mots-clés : TCP, la croissance de la fenêtre non-linéaire, les équations de Kolmogorov

1 Introduction

TCP congestion control is often analyzed using linear-increase multiplicative-decrease models for window variation [2, 8, 9, 13]. These models assume that the window increases linearly with time until a congestion occurs. At the moments of congestion, they assume that the window decreases multiplicatively by a factor of one half. The average round-trip time is used to calculate the window increase rate between congestion events. In particular, the window increase rate is taken equal to $1/(bRTT)$ packets/s, where b is the number of packets covered by a TCP acknowledgement (ACK) and RTT is the average round-trip time.

This simple model for window variation holds on long-distance paths where the throughput (that is, average transmission rate or the ratio of the total number of packets transmitted and the connection time) of a TCP connection is small compared to the total bandwidth. However, on short-distance paths where much bandwidth exists for each connection, two phenomena may appear making this model inaccurate. The first phenomenon is related to the receiver window. A TCP source cannot inject into the network in a single round-trip time more packets than the window advertised by the receiver [13, 14]. This puts a maximum limit on TCP window and, hence, makes an unlimited-window model overestimate the real performance.

The second phenomenon is related to the dependence between the window size and the round-trip time. When the share of a TCP connection from the total bandwidth is significant (due to a small number of concurrent connections), an increase in the window size very likely results in an increase in the round-trip time. The reason for this simultaneous increase is that at a large throughput, a TCP connection contributes considerably to the queueing time in network routers. An increase in the round-trip time together with an increase in the window size is known [1, 2, 4, 8] to result in a sublinear increase of the window size in time (the derivative of the window size with respect to time decreases). Hence, assuming that the window increases linearly with time while it increases sublinearly also results in an overestimation of the real performance [2].

We present a complete model for TCP congestion control. We account for both sublinearity and limitation of window increase. Some works in the literature account for such phenomena but they only consider simple networks of one bottleneck router and a single TCP connection [1, 4, 8]. In this paper, we consider real networks. To this end, we present a model for the variation of the round-trip time as a function of the window size. We propose a technique to infer the parameters of such model from the traces of a TCP connection. We then write the Kolmogorov equation of the window size in the stationary regime (we prove first the existence of such a regime) and we solve this equation numerically for the distribution of window size. The throughput of a TCP connection is computed from window size distribution. This throughput can be corrected for timeouts and the discrete nature of TCP congestion control using the heuristics in [2, 13].

In addition to the model for window variation, the modeling of TCP congestion control also requires a model for the moments at which the window is reduced [2]. We call these moments congestion moments or loss moments. First, we formulate the problem and we derive some stability results for any stationary and ergodic process of congestion moments. Then, we present a Markovian model for which we write the Kolmogorov equation of the window size distribution. This Markovian model is further specified in some particular cases, such as the cases of always-linear window growth and always-sublinear window growth.

The paper is organized as follows. In the next Section 2 we present the general model for the window size evolution as well as some stability results. The Markovian model is described in Section 3 with the help of its Kolmogorov differential equation. In Section 4 we present analytic solutions to the Kolmogorov equation for some particular cases. In Section 5 we show how to identify the parameters of the sublinear window size evolution, namely, we explain how to infer the parameters of the model for the round-trip time as a function of the window size. Finally, in Section 6 we present numerical and measurement results.

2 A general model for TCP

We first consider a very general model for the evolution of TCP congestion window. Our general model is composed of two parts: the model for window increase between loss moments and the model for loss moments. Recall that by a loss we mean an event that causes a reduction of TCP window.

Window evolution model between losses: Consider a fluid model of a TCP window [4, 8]. In the absence of losses, the window $W(t)$ (measured in packets) evolves according to

$$\frac{\mathrm{d}W}{\mathrm{d}t} = f(W), \tag{1}$$

where f is some nonnegative function (i.e., the window only decreases at the moments of congestion).

The main model that we shall analyze later will be the following special case of (1). It corresponds to the congestion avoidance mode of current TCP implementations [14]. In the absence of losses, the window W grows linearly with time until some threshold W_0 is achieved. Once the window size is greater than W_0, the growth becomes sublinear [1, 4, 8], and once a maximum window size M (determined by the receiver) is reached, the window remains constant at M. The sublinearity of window growth between W_0 and M is assumed to be caused by a linear increase in the round-trip time with the window size. Note that TCP sources increase their windows by the same amount of bytes in a round-time, whatever is the duration of the round-trip time. Hence, an increase in the round-trip time slows the window growth. Between 0 and W_0, the round-trip time is assumed to be constant and independent of the window size. The right hand side of the general model (1) is then given by

$$f(W) = 1\{W < M\}\frac{1}{bRTT(W)} = 1\{W < M\}\frac{1}{b(RTT_0 + \mu^{-1}1\{W > W_0\}(W - W_0))}, \tag{2}$$

where b is the number of data packets covered by an ACK (usually 2), RTT_0 is a *basic* round-trip time and the term $\mu^{-1}1\{W > W_0\}(W - W_0))$ corresponds to the increase in

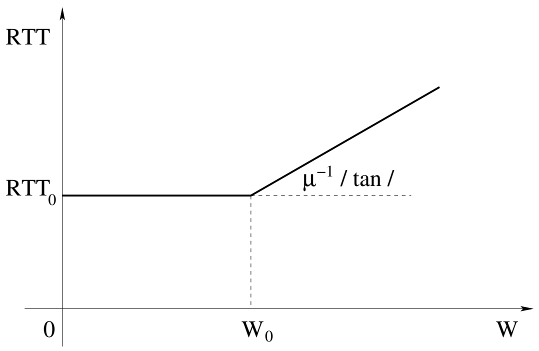

Figure 1: The dependence of round trip times on the window size

round-trip time caused by the queueing delay induced by the large window size. Figure 1 shows an example of how the round-trip time varies with the window size. For a simple network of one router and a single TCP connection [4, 8], RTT_0 represents the two-way propagation delay, μ represents the router bandwidth, and W_0 is simply equal to $RTT_0\mu$. In a real scenario, these quantities may have different interpretations. For example, the basic round-trip time can be seen as the sum of the propagation time and the contribution of the other flows to the queuing delay. We will propose in Section 5 to use the technique of non-linear least squares to infer these parameters from the traces of the TCP connection.

To see in more details how (2) is obtained, we recall that in the congestion avoidance mode, the window grows by approximately $1/W$ (packets) each time an ACK arrives. Let $ack(t)$ denote the number of arrivals of ACKs by time t. We can thus use our fluid approximation to express the rate of growth of the window by:

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \frac{\mathrm{d}W}{\mathrm{d}ack} \times \frac{\mathrm{d}ack}{\mathrm{d}t}; \tag{3}$$

When the window size reaches some value W_0, then the input rate of TCP reaches the capacity (bandwidth) of the bottleneck router, μ. At this point, packets leave the network at rate μ and enter at a larger rate. W_0 is given by $W_0 = RTT_0\mu$. When W_0 is reached, the excess is queued, so that the queue size at time t is $W(t) - W_0$, and the queueing time is $(W(t) - W_0)/\mu$. Thus the round trip time at time t is indeed $RTT_0 + (W(t) - W_0)/\mu$. At this point, the rate at which ACKs arrive is given by

$$\frac{\mathrm{d}ack}{\mathrm{d}t} = \frac{W(t)}{b \cdot RTT(t)} = \frac{W}{b(RTT_0 + \mu^{-1}1\{W > W_0\}(W - W_0))}.$$

Substituting this in (3) together with the rate of window growth (with respect to arrival of ACKs) finally yields (2).

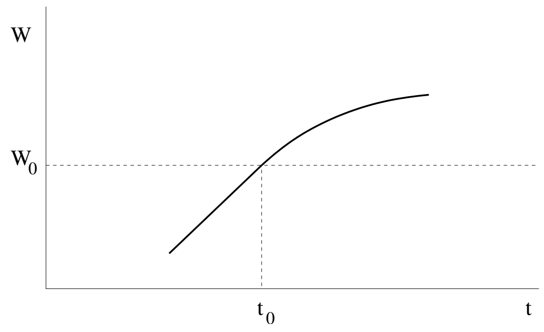

Figure 2: The sub-linear window evolution

Let t_0 denote the time when $W(t_0) = W_0$. Then, the window evolution for our main model is given by the function

$$W(t) = \begin{cases} W_0 + \frac{1}{bRTT_0}(t - t_0) & 0 < W \leq W_0, \\ \\ W_0 + \frac{\mu}{b}[\sqrt{b^2RTT_0^2 + 2\mu^{-1}b(t-t_0)} - bRTT_0] & W_0 < W < M, \end{cases}$$

see also Figure 2. We shall allow below as special cases to have $M = \infty$ (infinite receiver window) and/or $W_0 = \infty$ (always-linear growth) and/or $W_0 = 0$ (always-sublinear growth).

The loss model: We consider a stationary ergodic point process defined on some probability space $(\Omega, \mathcal{F}, P)$ with a finite rate $\nu > 0$, which will stand for the process of loss moments. We define a loss moment as the instant at which the window of a TCP connection is divided by a constant factor $\gamma > 1$. Typically $\gamma = 2$ [2, 13]. We consider a general reduction factor to account for other possible congestion control policies.

Let T_i, $i \in Z$, be the time instant at which the ith loss occurs, and denote by $\tau_i = T_{i+1} - T_i$ the ith inter-loss time. We take $... \leq T_{-1} \leq T_0 \leq 0 \leq T_1 \leq ....$ We shall allow in particular $\tau_i = 0$ with positive probability, which means that losses may arrive in batches.

We begin by establishing conditions for the tightness of the process W, and construct another simpler process that will serve as a majorant. This will allow us to obtain both bounds for performance measures as well as stability results.

Consider the process $\widehat{W}(t)$ which is defined on the same probability space as the original process W and is constructed as follows: it is also divided by γ at each loss, yet between losses it always grows linearly with some rate v, i.e.

$$\frac{\mathrm{d}\widehat{W}}{\mathrm{d}t} = v. \tag{4}$$

This process is well defined for any initial state. It has a unique stationary ergodic regime $\widehat{W}^*$, as was shown in [2]. In particular, if we consider the corresponding discrete time process embedded just prior to loss instants, then its unique stationary regime, which we denote by $\widehat{\mathcal{W}}_n^*$, is given by

$$\widehat{\mathcal{W}}_n^* = v \sum_{k=0}^{\infty} \gamma^{-k} \tau_{n-1-k}.$$

Lemma 1 *Consider a stationary ergodic loss point process with finite positive rate. Assume that there are two nonnegative numbers v and u such that $f(w) \leq v$ for any $w \geq u$. Then, for any fixed initial state at some time s which is taken to be the same in the original and new system (i.e., $W(s) = \widehat{W}(s)$), we have $\widehat{W}(t) + u \geq W(t)$ for all $t \geq s$.*

Note that the conditions of Lemma 1 apply in particular to our example (2) with $v = 1/(bRTT_0)$ and $u = 0$.

Define the correlation function between the inter-loss times: $R(k) = \mathrm{E}\left[\tau_n \tau_{n+k}\right]$. The above lemma implies that

Corollary 1 *Under the conditions of Lemma 1,*
(i) $W(t) \leq_{st} \widehat{W} + u$ for any t, and is therefore $W(t)$ is tight.
(ii) For any increasing function h:

$$\limsup_{t\to\infty} \mathrm{E}\left[h(W(t))\right] \leq \mathrm{E}\left[h(\widehat{W}^*(0))\right] + u$$

$$\limsup_{t\to\infty} \mathrm{E}\left[W(t)\right] \leq \lambda\alpha \left(\frac{1}{2} R(0) + \sum_{k=1}^{\infty} \gamma^{-k} R(k) \right) + u.$$

In the above corollary, "$\leq_{st}$" stands for the stochastic ordering (see e.g. [15]). The last equality follows from Proposition 2 in [2].

Remark 1 *Note that one can construct in a similar lower computable bounds for the process* $\mathrm{E}\left[h(W(t))\right]$ *if we replace the condition $f(w) \leq v$ for $w \geq u$ in Lemma 1 by $f(w) \geq v$.*

In the next two theorems we provide two stability results. The first establishes the existence of a stationary ergodic regime, whereas the second one establishes its uniqueness and convergence of the window size to that regime.

Theorem 1 *Assume that the loss process is stationary ergodic. Assume further that there are two nonnegative numbers v and u such that $f(w) \leq v$ for any $w \geq u$. Then there exists a stationary ergodic process $W^*(t)$ satisfying the evolution (1) between losses and for which $W(t)$ is divided by a factor of γ for each loss.*

PROOF: Define on the same probability space the family of processes $\{W^{(s)}(t), t \in R\}$, $n \in R$ as follows. $W^{(s)}(t) := 0$ for $t \leq -s$, and for $t > -s$ it is given by the TCP evolution described by (1) and with the window divided by γ for each loss. Thus all the processes $W^{(s)}$ experience losses at the same instants. For each t, $W^{(s)}(t)$ is increasing with respect to s and thus it has a limit $W^*(t)$. The limit is clearly finite if $M < \infty$. Next we show that in the case of $M = \infty$, the limit is finite almost surely.

Consider the process $\widehat{W}(t)$ defined on the same probability space defined in (4). Let $\widehat{W}^*$ be the unique stationary regime corresponding to $\widehat{W}$ (see [2]). It follows from Lemma 1 that our limit process W^* is majorized by the stationary ergodic process $\widehat{W}^* + u$, and therefore it is finite a.s. (and tight). Since W^* is a function of the stationary ergodic loss process, it is also ergodic. This establishes the theorem.

□

We call the process W^*, defined in the previous theorem, the *minimal stationary regime* of W.

Theorem 2 *Assume that the loss process is stationary ergodic. Assume that $f(W)$ is non-increasing. Then there is a unique stationary regime W^* and for any initial value $W(0)$, we have*

$$\lim_{t\to\infty} |W(t) - W^*(t)| = 0$$

almost surely.

We shall use the following obvious observation:

Lemma 2 *Let $W(t, w_0)$ be the process $W(t)$ obtained when starting initially at $W(0) = w_0$. Then, for any $t \geq 0$, $W(t, w_0)$ is monotone in w_0.*

PROOF OF THEOREM 2: Define $S_n = T_n$ for $n > 0$ and $S_0 = 0$. Consider the embedded process $W_n := W(S_n)$ for all nonnegative integers, with some initial condition $W_0 = W(0)$. Consider on the same probability space another embedded process $W'_n := W'(T_n)$ which is obtained similarly using the same dynamics as that defining W, but with an initial condition $W'_0 > W_0$. It follows from Lemma 2 that $W'_n \geq W_n$ for all positive integers n. Now,

$$\begin{aligned} W'_{n+1} - W_{n+1} &= \gamma^{-1}\Big(W'_n - W_n + \int_{S_n}^{S_{n+1}} \Big[\frac{dW'(t)}{dt} - \frac{dW(t)}{dt}\Big] dt\Big) \\ &= \gamma^{-1}\Big(W'_n - W_n + \int_{S_n}^{S_{n+1}} (f(W'(t)) - f(W(t)))dt\Big) \\ &\leq \gamma^{-1}(W'_n - W_n) \end{aligned}$$

where we used the fact that f is non-increasing and Lemma 2. It follows that

$$W'_{n+1} - W_{n+1} \leq \gamma^{-n}(W'_0 - W_0),$$

which shows that the processes W_n converges to a limit regime that does not depend on the initial state. Since T_i are finite P a.s. (as the loss point process is assumed to have a positive rate), we conclude that $W(t)$ also converges to a unique limiting process that does not depend on the initial state. This proves the theorem.

$\square$

Remark 2 *One can easily show that in general, when f does not satisfy the conditions of Theorem 2, there need not be a unique stationary regime for the process W. Consider, for example, the case where*

$$\gamma = 2, \qquad f(w) = 1 + 9 \times 1\{w > 5\}.$$

First we suppose that losses occur just after $T_n = n$. If $W(0) = 0$, the window process converges to

$$W^*(t) = t - \lfloor t \rfloor + 1, \quad t \in \mathbb{R}$$

where $\lfloor t \rfloor$ denotes the largest integer number which is smaller than t. This process takes values in the interval $[1, 2]$. However, if the process starts at sufficiently large $W(0)$, it converges to another limit process:

$$\overline{W}^*(t) = 10W^*(t),$$

which takes values in $[10, 20]$. Although in this example the limit processes are not stationary ergodic, by adding slight perturbation to the time between losses (e.g., by letting τ_n be i.i.d. random variable, uniformly distributed in $[1, 1 + \epsilon]$ for some ϵ small), we obtain the same features as in the above example yet with the limiting processes being stationary [5].

3 A Markovian model

In this section we study the window evolution according to the dynamics described by (2). For the loss process we consider the following Markovian batch model.

We assume that batches containing a random number of losses arrive according to an independent Poisson process with intensity λ. The window is divided by a factor $\gamma > 1$ for each loss in a batch. We denote the sizes of (i.e., the numbers of losses in) consecutive batches by $N_1, N_2, N_3, \ldots$, and we assume that these constitute an i.i.d. sequence. The size of an arbitrary batch is generically denoted by $N \stackrel{d}{=} N_k$. The Poisson process and the sequence N_k, $k = 1, 2, \ldots$, are independent of each other and independent of the past evolution of the window. For a batch containing n losses, the window is multiplicatively decreased by a factor γ^{-n}. Immediately after the multiplicative decrease, the window restarts its growth (linear or sublinear). Furthermore, the window stays constant at M when this maximum level is reached until the next batch of losses appears.

Note that $W(t)$ is a so-called "Piecewise Deterministic Markov Process", see [6]. We denote the probability generating function of the distribution of N by

$$Q(z) := \mathrm{E}\left[z^N\right] =: \sum_{n=1}^{\infty} z^n q_n, \quad |z| \le 1. \tag{5}$$

We are interested in the calculation of the stationary distribution function of $W(t)$, that is

$$F(x) := \lim_{T\to\infty} \frac{1}{T} \int_{t=0}^{T} \mathbf{P}\left\{W(t) \le x\right\} \mathrm{d}t. \tag{6}$$

Once this distribution is calculated, the throughput (in packets/s) can be deduced in the following way [2]

$$\bar{X} = \mathrm{E}\left[\frac{W^*}{RTT(W^*)}\right] = \int_0^{M+} \frac{x}{RTT(x)} \mathrm{d}F(x).$$

Note that the throughput is no other than the expectation of the instantaneous transmission rate $X(t) = W(t)/RTT(W(t))$. To correct for the burstiness of TCP, the instantaneous transmission rate is supposed to be averaged over the last round-trip time.

The next theorem states that the distribution $F(x)$ exists and is unique. It also provides the Kolmogorov steady-state differential equations.

Theorem 3 *There exists a unique steady-state distribution of the window size for the window evolution model defined in* (2) *and the batch loss Poisson process. The complementary distribution function* $\overline{F}(x) = 1 - F(x) = \boldsymbol{P}\{W > x\}, x \in (0, M]$, *is a solution of the following Kolmogorov steady-state differential equations*

$$-\frac{1}{b(RTT_0 + \mu^{-1}1\{x > W_0\}(x - W_0))} \frac{\mathrm{d}}{\mathrm{d}x}\overline{F}(x) = \lambda\left(\overline{F}(x) - \sum_{n=1}^{\infty} q_n \overline{F}(\min(\gamma^n x, M))\right), \tag{7}$$

where $x \in (0, M)\backslash\{M/\gamma^n\}_{n=1,2,\ldots}$.

Proof: The existence and uniqueness of the steady-state distribution follows immediately from Theorem 2, as the function $f(W)$ defined in (2) is indeed non-increasing and the batch loss Poisson process is stationary and ergodic. To derive the Kolmogorov steady-state differential equation we use the up and down crossing argument. Namely, assume that the process is in equilibrium and consider a level $x \in (0, M)$. Whenever the window size increases from less than or equal to x to more than x we say that an up crossing of the level x has occurred. Similarly, if the window size decreases from more than x to less than or equal to x we say that a down crossing of the level x has occurred. Let $[t, t + \Delta]$ be a small deterministic time interval. When the process is in equilibrium, the probability of up-crossing

$$(1 - \lambda\Delta)\mathbf{P}\left\{x - \frac{1}{b(RTT_0 + \mu^{-1}1\{x > W_0\}(x - W_0))}\Delta < W \le x\right\} + \mathrm{o}(\Delta)$$

is equal to the probability of down-crossing

$$\lambda\Delta \sum_{n=1}^{\infty} q_n \mathbf{P}\left\{x < W \le \min(\gamma^n x, M)\right\} + \mathrm{o}(\Delta).$$

After equating these probabilities, we pass $\Delta \downarrow 0$. We note that the derivative of $F(x)$ exists and is continuous for all x except at $x = W_0$ and $x = M\gamma^{-n}$, when $q_n > 0$. For $x \in (0, M)\backslash\{M\gamma^{-n}\}_{n=1,2,...}$ we obtain the following steady-state Kolmogorov equation

$$\frac{1}{b(RTT_0 + \mu^{-1}1\{x > W_0\}(x - W_0))} \frac{\mathrm{d}}{\mathrm{d}x}\mathbf{P}\{W \le x\} = \lambda \sum_{n=1}^{\infty} q_n \mathbf{P}\{x < W \le \min(\gamma^n x, M)\}$$

The above equation immediately imply (7).

□

The differential equation (7) can be solved in a recurrent manner. Namely, we solve the equation successively on the intervals $[M/\gamma, M)$, $[M/\gamma^2, M/\gamma]$,... Note that at each step of this recursion one needs to solve a linear differential equation of the first order. Thus, in principle, one can obtain an analytic solution for any number of the intervals $[M/\gamma^{n+1}, M/\gamma^n]$. However, the analytic solution is very messy and it is recommended to use any standard numerical differential equation solver. The probability $P_M = \mathbf{P}\{W^* = M\}$ can be found from the normalization condition. Furthermore, P_M as well as the moments of the window size distribution can be obtained explicitly in some particular cases (see Section 4).

4 Some important particular cases

In this section we present some particular but important cases when we can obtain simple analytic expressions for the distribution and the moments of the window size as well as for the constant P_M.

4.1 The case of only linear window growth

Here we present the results for the case $W_0 \ge M$, that is, we assume that the window growth is always linear. The linear window growth holds on paths where the connection under consideration does not contribute significantly to queueing delays. For the detail derivations we refer an interested reader to [3]. If $W_0 \ge M$, the coefficient in front of the derivative in (7) becomes constant $\alpha := 1/(bRTT_0)$. Consequently, on the interval $[M/\gamma, M)$ the distribution function is given by

$$\overline{F}(x) = P_M \mathrm{e}^{\frac{\lambda}{\alpha}(M-x)}, \quad \frac{M}{\gamma} \le x < M.$$

We recall that $P_M = \mathbf{P}\{W^* = M\}$. Let $\overline{F}_k(x)$ denote the truncation of $\overline{F}(x)$ on the interval $[M/\gamma^k, M/\gamma^{k-1})$. Then,

$$\overline{F}_k(x) = P_M \sum_{i=1}^{k} c_i^{(k)} \mathrm{e}^{-\frac{\lambda}{\alpha}\gamma^{i-1}x}, \quad k = 1, 2, ... \tag{8}$$

The coefficients $c_i^{(k)}$ are calculated by the following double recursion on k and i

$$c_{i+1}^{(k)} = \frac{1}{1-\gamma^i} \sum_{n=1}^{i} q_n c_{i-n+1}^{(k-n)}, \quad i = 1, ..., k-1,$$

with $c_1^{(k)}$ given by

$$c_1^{(k)} = e^{\frac{\lambda}{\alpha}\frac{M}{\gamma^{k-1}}} \left[\sum_{i=1}^{k-1} c_i^{(k-1)} e^{-\frac{\lambda}{\alpha}\gamma^{i-k}M} - \sum_{i=2}^{k} c_i^{(k)} e^{-\frac{\lambda}{\alpha}\gamma^{i-k}M} \right].$$

The probability P_M of the window size being at the maximum level can be calculated by the formula

$$P_M = \left(1 + \left(1 - Q(\gamma^{-1})\right) \sum_{i=0}^{\infty} e_i \left(\mathrm{e}^{\gamma^{-i}\frac{\lambda}{\alpha}M} - 1 \right) \right)^{-1},$$

where

$$\frac{e_i}{e_0} = \frac{1}{1-\gamma^{-i}} \sum_{n=1}^{i} \gamma^{-n} q_n \frac{e_{i-n}}{e_0}, \quad i \geq 1,$$

$$e_0 = \left(1 + \sum_{n=1}^{\infty} \gamma^{-n} q_n \sum_{j=0}^{\infty} \frac{e_j/e_0}{\gamma^{j+n} - 1} \right)^{-1}.$$

Next, define for $\mathrm{Re}(\omega) \geq 0$ the LST (Laplace-Stieltjes Transform) of the window size distribution by

$$\hat{f}(\omega) = \int_{x=0}^{M+} \mathrm{e}^{-\omega x} \mathrm{d}F(x).$$

Taking Laplace Transforms in (7) leads to:

$$\alpha \left(\hat{f}(\omega) - P_M \mathrm{e}^{-\omega M} \right) = \lambda \frac{1 - \hat{f}(\omega)}{\omega} - \lambda \sum_{n=1}^{\infty} \gamma^{-n} q_n \frac{1 - \hat{f}(\gamma^{-n}\omega)}{\gamma^{-n}\omega}. \tag{9}$$

Since $\mathrm{E}\left[W^k\right] \leq M^k$, $k = 1, 2, \ldots$, we may write

$$\hat{f}(\omega) = 1 + \sum_{k=1}^{\infty} \frac{(-\omega)^k}{k!} \mathrm{E}\left[W^k\right].$$

Substituting the above series in (9), using the absolute convergence of the doubly-infinite series to interchange the order of summation and equating the coefficients of equal powers of ω we get the following recursive formula for the moments of the window size distribution

$$\mathrm{E}\left[W^k\right] = \frac{k\alpha\left(\mathrm{E}\left[W^{k-1}\right] - P_M M^{k-1}\right)}{\lambda\left(1 - Q(\gamma^{-k})\right)}, \quad k = 1, 2, \ldots$$

In particular we find for $k = 1, 2$:

$$\mathrm{E}\left[W\right] = \frac{\alpha\left(1 - P_M\right)}{\lambda\left(1 - Q(\gamma^{-1})\right)},$$

$$\mathrm{E}\left[W^2\right] = \frac{2\alpha\left[\alpha\left(1 - P_M\right) - \lambda P_M M\left(1 - Q(\gamma^{-1})\right)\right]}{\lambda^2\left(1 - Q(\gamma^{-1})\right)\left(1 - Q(\gamma^{-2})\right)}.$$

The round-trip time is constant in this case (linear window growth) and the throughput is simply equal to $\bar{X} = \mathrm{E}\left[W\right]/RTT_0$.

4.2 The case of only sublinear window growth

Here we study the cases when the round-trip time always grows linearly with the window size. This corresponds to congested paths where a queue of packets always exists in routers. A path crossed by TCP connections is congested either because the bandwidth-delay product is small compared to the buffering capacity of routers, or because the number of connections is large [8]. First, we consider the case when the constant component of the round-trip time (RTT_0) is negligible compared to the increase in round-trip time induced by the connection. In our model such a situation would correspond to $RTT_0 \approx 0$ and $W_0 \approx 0$. If we assume $RTT_0 = W_0 = 0$, the differential equation for the window size evolution between losses takes the following form

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \frac{\mu}{bW}. \tag{10}$$

The window reduction at instants of losses is as before and $W(t)$ stays constant until the next loss when the maximum window size M is reached. As before, we seek to find a stationary probability distribution for $W(t)$ that satisfies these dynamics. Our approach will be to transform (10) into an equation of the type studied in Section 4.1. This can be achieved using the transformation[1] $X(t) = W(t)^2$, which indeed leads to

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \frac{2\mu}{b},$$

i.e., a constant linear growth in between loss instants. The maximum value of the transformed process $X(t)$ is, of course, M^2. If at the k-th loss instant $t = T_k$, the window $W(T_k^-)$ is reduced by a factor γ^{n_k} due to n_k clustered loss events, then

$$X(T_k^+) = W(T_k^+)^2 = \left(\gamma^{-n_k} W(T_k^-)\right)^2 = \left(\gamma^2\right)^{-n_k} X(T_k^-),$$

that is, the value of the process $X(t)$ is reduced by a factor γ^2 (instead of γ) for each individual loss event. We can compute the stationary distribution of $X(t)$ as in Section 4.1

[1] The transformation $X(t) = W(t)^{m+1}/(m+1)$ has been used in Ott et al. [12] to analyze the more general case $\mathrm{d}W/\mathrm{d}t = cW^{-m}$, $c > 0, m > -1$ for single losses ($q_1 = 1$) and unlimited window growth ($M = \infty$).

taking a maximum level M^2 (instead of M), a linear increase rate $\alpha = 2\mu/b$ and a reduction factor γ^2 (instead of γ). With these substitutions, the complementary distribution function of $X(t)$ is given by (8) and, hence, the stationary version W of the process $W(t)$ satisfies:

$$\mathbf{P}\{W > x\} = P_M \sum_{i=1}^{k} c_i^{(k)} \mathrm{e}^{-\frac{\lambda}{\alpha}\gamma^{2(i-1)}x^2}, \quad x \in [M/\gamma^k, M/\gamma^{k-1}),\ k = 1, 2, \ldots \tag{11}$$

The probability $P_M := \mathbf{P}\{W = M\} = \mathbf{P}\left\{X = M^2\right\}$ and the coefficients $c_i^{(k)}$ are calculated as in Section 4.1 (with M, α and γ replaced by M^2, $2\mu/b$ and γ^2, respectively). The throughput in this case is simply equal to μ. Note that from the results of Section 4.1 we can also immediately obtain a recursion on the even moments of the window size distribution. In the following we show, however, how such a recursion can also be obtained for the odd moments, in the more general case where $RTT_0 > 0$.

Next, we consider the case when the constant component RTT_0 of the round-trip time is significant. W_0 is still assumed to be negligible due to a persistent congestion of the path. This leads to the following Kolmogorov equation for the steady-state window size distribution

$$-\frac{1}{b(RTT_0 + x/\mu)}\frac{\mathrm{d}}{\mathrm{d}x}\overline{F}(x) = \lambda\left(\overline{F}(x) - \sum_{n=1}^{\infty} q_n \overline{F}(\min(\gamma^n x, M))\right).$$

By multiplying the above equation by $x^k b(RTT_0 + x/\mu)$ and then integrating from 0 to M^-, we obtain the next recurrent relation between the moments $W^{(k)} = \mathrm{E}\left[W^k\right], k \geq 1$

$$W^{(k)} - M^k P_M = \frac{\lambda b RTT_0}{k+1}[1 - Q(\gamma^{-(k+1)})]W^{(k+1)} + \frac{\lambda b \mu^{-1}}{k+2}[1 - Q(\gamma^{-(k+2)})]W^{(k+2)}.$$

If the first moment $W^{(1)} = \mathrm{E}[W]$ and the constant P_M are determined (e.g., after having numerically determined the distribution function), then the higher moments can be calculated by the simple recurrent formula:

$$\begin{aligned} W^{(k+2)} &= -\frac{\mu RTT_0(k+2)[1 - Q(\gamma^{-(k+1)})]}{(k+1)[1 - Q(\gamma^{-(k+2)})]}W^{(k+1)} \\ &\quad + \frac{(k+2)}{\lambda b \mu^{-1}[1 - Q(\gamma^{-(k+2)})]}W^{(k)} - \frac{(k+2)M^k}{\lambda b \mu^{-1}[1 - Q(\gamma^{-(k+2)})]}P_M, \quad k \geq 2. \end{aligned}$$

5 Identification of round-trip time model parameters

For networks of one bottleneck router and a single TCP connection (see [1, 4, 8]), the three parameters of the model for round trip time (RTT_0, μ and W_0) can be directly deduced. For more complicated networks, these parameters have to be inferred on end-to-end basis from the traces of the connection. Assume that we have some measurements of the round

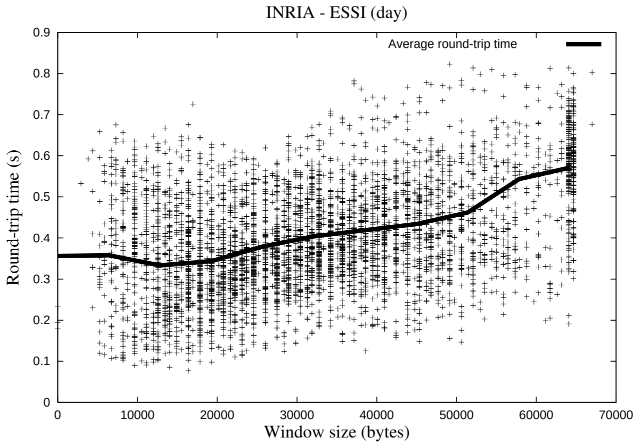

Figure 3: Round-trip time vs. window size

trip times and the corresponding window sizes seen by the connection. Figure 3 shows an example of such measurements for a TCP connection that we ran for twenty minutes between a machine at INRIA Sophia Antipolis (south of France) and another machine at ESSI about 1 Km from INRIA location. Each point in this figure represents a measurement of the round trip time and the corresponding window size. These points are obtained by a tool that we developed and that monitors the flow of packets and ACKs at the output interface of the TCP source machine. The thick line in the figure represents the average round trip time over close windows. It is clear how the round trip time tends to increase with the window size. Next, we explain how we can infer the three parameters of our model from such traces.

Let RTT_i be the i-th measurement of the round-trip time and let W_i be the corresponding window size. When using our model to predict the round-trip time for the window size W_i, the error we introduce is equal to

$$\epsilon_i = RTT_0 + \mu^{-1} 1\{W_i > W_0\}(W_i - W_0) - RTT_i.$$

Let n be the total number of measurements. We propose to use the non-linear least-square technique which consists in finding the parameters of the model that minimize the sum $\sum_{i=1}^{n} \epsilon_i^2$. We solve numerically such minimization problem for the traces presented in Figure 3. The program in C that we developed using the non-linear simplex method of the NAG library [11] gives the curve shown in Figure 4. The figure also shows 95% confidence intervals.

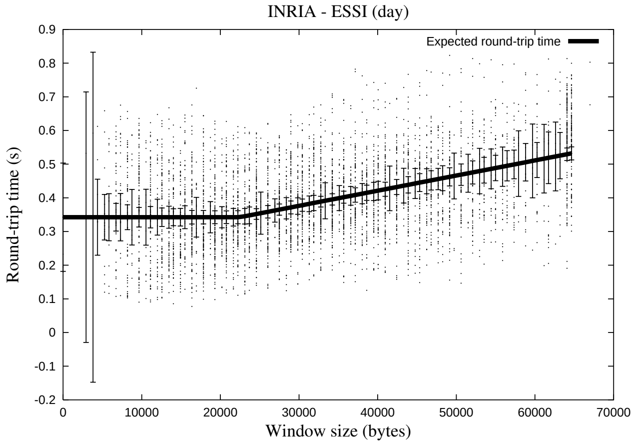

Figure 4: Expected round-trip time vs. window size

6 Measurement results

We use the traces of the TCP connection between INRIA and ESSI to validate our calculation of the throughput and the window size distribution. This connection was run for a whole day in January 2000. We only consider the working hours. Approximately every twenty minutes, we store the traces of the connection in separate files. Then, we apply our Markovian model to predict the throughput of the connection in each time interval. We use two variations of our model. First, we use the model for the case of always-linear window growth. Then, we use the general model which takes into account the both cases of linear and sublinear window growth (see equation (7)). The non-linear least-square technique is applied to the traces for different time intervals to find the parameters of the model for round trip time variation. In both cases, we take $N \equiv 1$, that is, the moments at which the window is reduced are assumed to follow a Poisson process. The maximum window size on this connection is equal to 64 Kbytes and the New-Reno version of TCP is used in the source machine at INRIA [7].

First, we plot in Figure 5 the variation of the throughput of the connection during the day corresponding to the both models we considered. We also plot the variation of the real throughput. The linear model overestimates the real throughput on this connection, whereas the estimation given by the general model is much more accurate. The overestimation given by the linear model is the result of the sublinearity of the window increase on this short-distance path. The model of round trip time for this connection shown in Figure 4 is a clear proof of such sub-linearity.

Second, we plot the window size distribution function $F(x)$ at different hours during the day. Two samples are shown in Figure 6. The figure shows the distribution function given by our linear and sublinear models as well as that calculated from window size measurements.

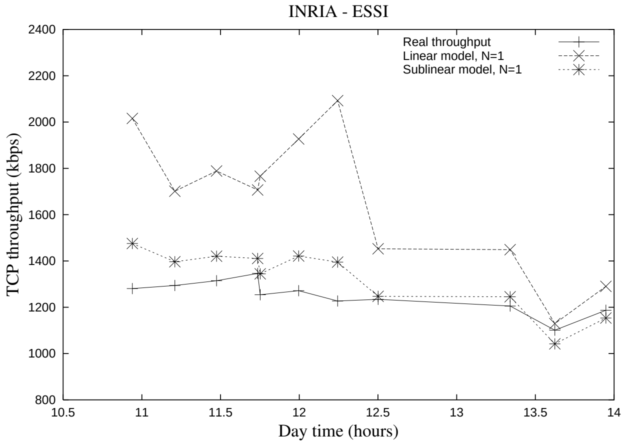

Figure 5: TCP throughput: sublinear vs. linear models

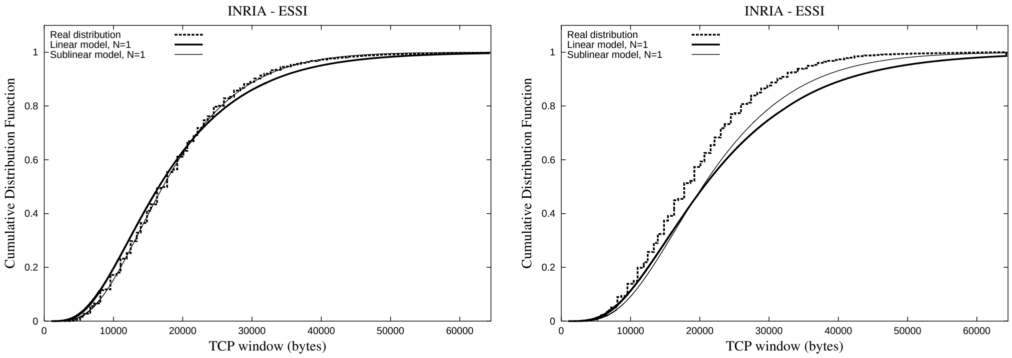

Figure 6: Window size distribution function

A good match is noticed between the model and the reality. The figures (especially the right-hand one) also shows that the linear model overestimates the real throughput by giving more weight to large windows.

Contents

1 Introduction **3**

2 A general model for TCP **4**

3 A Markovian model **9**

4 Some important particular cases **11**
4.1 The case of only linear window growth 11
4.2 The case of only sublinear window growth 13

5 Identification of round-trip time model parameters **14**

6 Measurement results **16**

References

[1] A.A. Abouzeid, S. Roy, and M. Azizoglu, "Stochastic modeling of TCP over lossy link", IEEE INFOCOM'2000.

[2] E. Altman, K. Avrachenkov and C. Barakat, "A stochastic model of TCP/IP with stationary random losses", *ACM Sigcomm*, Aug. 28 - Sept. 1, Stockholm, Sweden, 2000.

[3] E.Altman, K.Avrachenkov, C.Barakat, and R. Núñez-Queija, "State-dependent M/G/1 type queueing analysis for congestion control in data networks", IEEE INFOCOM'2001.

[4] E. Altman, J. Bolot, P. Nain, D. Elouadghiri- M. Erramdani, P. Brown, and D. Collange, "Performance Modeling of TCP/IP in a Wide-Area Network", *INRIA Research Report N=3142*. (available in http://www.inria.fr:80/RRRT/RR-3142.html). A shorter version in: *34th IEEE Conference on Decision and Control*, Dec 1995.

[5] S. Asmussen, "Applied Probability and Queues", Wiley, 1987.

[6] M. H. A. Davis, *Markov Models and Optimization*, Chapman and Hall, London, Glasgow, N.Y., Tokyo, Melbourne, Madras, 1993.

[7] S. Floyd and T. Henderson, "The NewReno Modification to TCP's Fast Recovery Algorithm", *RFC 2582*, Apr. 1999.

[8] T. Lakshman and U. Madhow, "The performance of TCP/IP for networks with high bandwidth-delay products and random loss", IEEE/ACM Transactions on Networking, v.5, no. 3, 1997.

[9] M. Mathis, J. Semke, J. Mahdavi, and T. Ott, "The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm", *ACM Computer Communication Review*, vol. 27, no. 3, pp. 67-82, Jul. 1997.

[10] V. Misra, W.-B. Gong, and D. Towsley, "Stochastic differential equation modeling and analysis of TCP-windowsize behaviour", *Performance*, Oct. 1999.

[11] Mathematical C library, http://www.nag.co.uk/

[12] T.J. Ott, J.H.B. Kemperman, and M. Mathis, "The stationary behavior of ideal TCP congestion avoidance", available at: ftp://ftp.telecordia.com/pub/tjo/TCPwindow.ps, Aug. 1996.

[13] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, "Modeling TCP Throughput: a Simple Model and its Empirical Validation", *ACM SIGCOMM*, Aug. 1998.

[14] W. Stevens, "TCP Slow-Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms", *RFC 2001*, Jan. 1997.

[15] V. Strassen, "The existence of Probability Measures with Given Marginals," *Ann. Math. Statist.* **36**, pp. 423-439, 1965.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399