A Stochastic Model of TCP/IP with Stationary Random Losses*

Eitan Altman, Konstantin Avrachenkov†, Chadi Barakat‡

INRIA, 2004 route des Lucioles, 06902 Sophia Antipolis, France
Email:{altman,kavratch,cbarakat}@sophia.inria.fr

***Abstract*— In this paper, we present a model for TCP/IP congestion control mechanism. The rate at which data is transmitted increases linearly in time until a packet loss is detected. At this point, the transmission rate is divided by a constant factor. Losses are generated by some exogenous random process which is assumed to be stationary ergodic. This allows us to account for any correlation and any distribution of inter-loss times. We obtain an explicit expression for the throughput of a TCP connection and bounds on the throughput when there is a limit on the window size. In addition, we study the effect of the Timeout mechanism on the throughput. A set of experiments is conducted over the real Internet and a comparison is provided with other models that make simple assumptions on the inter-loss time process. The comparison shows that our model approximates well the throughput of TCP for many distributions of inter-loss times.**

I. Introduction

We analyze in this paper the performance of TCP (Transmission Control Protocol), the widely-used transport protocol of the Internet [21], [38]. TCP is a reliable window-based flow control protocol where the window is increased until a packet loss is detected. Here, the source assumes that the network is congested and reduces its window. Once the lost packets are recovered, the source resumes its window increase. As a performance measure, we consider the throughput of a long-lived TCP connection having an infinite amount of data to send. A mathematical model is presented to find a closed-form expression for the throughput of the connection.

We assume that the reader is familiar with basic mechanisms of TCP [38] such as Slow Start and Congestion Avoidance algorithms, the two methods for loss detection: Duplicate ACKs and Timeout, the Delay ACK mechanism, the limitation on the congestion window due to receiver or sender buffer, etc. (see [8] for a survey on TCP issues).

A remarkable attention has been given to TCP modeling within the research community, e.g., [2], [15], [23], [24], [28], [29], [30], [35], [37]. This is not surprising since 95% of Internet traffic is carried over TCP [39]. Closed-form expressions for the throughput of a long-lived TCP connection have been obtained under different assumptions. These expressions have helped to understand the impact of network and TCP parameters on the throughput of the connection and on the efficiency of network resource utilization. Recently, these expressions have been also used to adapt the rate of UDP flows (e.g., audio and video) in a way to be friendly with TCP flows [16].

The mathematical analysis of TCP requires two steps. First, we need to construct a model for the window size evolution. Since most of Internet traffic in terms of bytes is carried by long-lived TCP connections, the majority of existing models focus on the Congestion Avoidance mode. A fluid model is often used. The window of TCP is assumed to increase linearly as a function of time until a loss occurs, and it is divided by two when the loss is detected. An initialization to one packet is proposed in [30] for losses detected via Timeout. The phase of recovery from losses is assumed to be negligible and the source is assumed to resume the linear increase of its congestion window directly after the reduction. In [35], a packet-level model is proposed to account for the discrete nature of TCP. Indeed, the volume of data in the network is at any moment in multiple of packets (packet size equal to MSS – Maximum Segment Size) due to the Nagle algorithm [31], which for efficiency reasons, prohibits TCP from injecting into the network packets of small size (smaller than MSS). This volume of data increases by one packet when the increase in the window size exceeds the packet size. Later in our paper, we will show how a fluid model can be corrected to account for this discreteness of TCP.

Second, TCP analysis requires a characterization of times between congestion events. Namely, one needs to model the impact of the path between the source and the destination on the TCP connection. Particular models are considered in the literature. The fixed point approach used in [24], [28] assumes a constant time between congestion events. The assumptions made in [35] can also be shown to imply a constant time between congestion events. In [30], congestion events are modeled by a homogenous Poisson process. In [37], the intensity of the Poisson process is assumed to increase with the window size. Instead of working in real time, the authors in [29], [34] chose to work in a virtual time, which is obtained by sampling the congestion window of TCP at the moments of ACK arrivals. They consider the case where times between congestion events in this virtual time are identically and exponentially distributed. The distribution as well as the moments of the congestion window size are found in this virtual time and a method is suggested to transform them back

* An earlier version of this paper has appeared in ACM SIGCOMM 2000.
† The work of this author was financed by a grant of CNET France-Telecom on flow control in High Speed Networks.
‡ The work of this author was financed by an RNRT “Constellations” project on satellite communications.
‡ Contact author.

to the real time.

Our experimentations over the Internet show that times between congestion events can have general distributions. Depending on the monitored path, these times can vary from an approximately deterministic case to a considerably bursty case. Moreover, some correlation can exist between congestion events. We note in particular that if packets are dropped independently of each other with constant probability, then the times between drops are not independent (since the instantaneous transmission rate is variable). We believe that the Internet is so heterogeneous that different types of distributions of times between congestion events will always exist.

In this paper, we investigate the case of a general sequence of times between congestion events. In the sequel, we will call a congestion event a *loss event*. A loss (event) corresponds to a moment where the congestion window is divided by a constant factor – usually equal to 2.[1] We only require that this process of loss events is stationary ergodic. With this minimal requirement we are able to obtain an explicit expression for the throughput of TCP. Our loss model is general enough to capture any correlation and any distribution of inter-loss times.

Obtaining a closed-form expression of TCP throughput for general loss processes can be quite useful for many design and dimensioning purposes. It could be used for the fine-tuning of physical transmission channels; for example on satellite links, coding schemes that include redundancy and interleaving cause losses to appear in bursts (see [20]). Formulas that take into account this burstiness can help to predict the impact of coding schemes on TCP connections, and hence to optimize the amount of redundancy that should be added. A model is presented in [9] to optimize the amount of redundancy on a noisy link using a formula for TCP throughput. Another important use of closed-form expressions for the average transmission rate (or equivalently the throughput) is in the design of TCP-friendly applications. The latter are typically real time applications that are designed to use a fair share of the bandwidth in comparison with TCP connections (see [16], [17] and references therein). As we will see, when loss events are highly bursty, the transmission rate computed under the assumption of deterministic or exponential inter-loss times, considerably underestimate the throughput of a TCP connection. Hence, TCP-friendly applications can perform better and transmit at a higher rate when using a more precise model for the losses.

As for the dynamics of TCP, we model the instantaneous transmission rate which is defined as the number of packets in the network (or the volume of data) divided by the RTT (Round-Trip Time) of the connection. The TCP source is assumed to always have data to send. The number of packets in the network is thus equal to the number of packets that can fit within the window. Denote by $X(t)$ the transmission rate of the TCP connection at time t averaged over RTT.[2] We assume that $X(t)$ increases linearly with time at a rate α.[3] If we denote by b the number of data packets covered by one ACK and by RTT the average round-trip time, we find $\alpha = 1/(bRTT^2)$. Let ν denote the decrease in the transmission rate when a loss event occurs.[4] The arrivals of losses are modeled by a general stationary ergodic point process [7] with non-null and finite intensity λ. Let $\{T_n\}_{n=-\infty}^{+\infty}$ be a particular realization of the point process. Consider for instance the case when losses are quickly detected without the need for a long Timeout period (e.g. via the three duplicate ACKs algorithm or an efficient fine-granularity Timeout mechanism). Then, the evolution of the transmission rate can be described by the following recurrence

$$X_{n+1} = \nu X_n + \alpha S_n, \tag{1}$$

where X_n is the value of $X(t)$ just prior to the arrival of the loss at T_n, and $S_n := T_{n+1} - T_n$. The pair $\{T_n, X_n\}$ can be considered as a marked point process [7]. As we will see, the model that we consider here allows in particular for the distribution of S_n to depend on X_n.

In the next section we use the machinery of stochastic processes to study this model of TCP rate evolution. We first introduce some tools to handle (1), and in particular to handle the case where the distribution of S_n may depend on X_n. Using these tools we compute the throughput, i.e. the time average of process $X(t)$. We also compute the first two moments of the TCP transmission rate at loss arrivals for the stationary regime. The model can be used to compute all the higher moments of the transmission rate of TCP in the stationary regime. In particular, we will show how to find the variance of $X(t)$. Then different examples of loss processes are studied: Deterministic, Poisson, i.i.d. and Markov arrival processes. The expression of the throughput is provided for each of these particular cases. In Section II-E, we extend our model to account for the case when there is a limitation on the evolution of the transmission rate (e.g. due to the receiver advertised window); we provide bounds on the throughput for this case. In Section II-F, we explain how to extend our model to the case when some losses are detected via a conservative coarse-granularity Timeout mechanism, which is used in most TCP implementations. In Section III, we present the testbed as well as the results of our experimentations. The results demonstrate that different types of loss processes exist in the Internet, and that often the distribution of inter-loss times cannot be approximated by a constant or by the exponential distribution. The experimentations also show the common problem of linear rate increase models. On traces where the transmission rate of TCP exhibits a linear increase, our model gives excellent results. However, on traces where the TCP window grows sub-linearly, linear models overestimates the real throughput. We conclude Section III with a method to correct the error caused by the fluid approximation. Finally, we present our conclusions in Section IV.

[1] This division can be the result of multiple packet losses. Ideally, a TCP connection must divide its window by two whatever is the number of packet losses within a Round Trip Time (RTT) [35].

[2] At any time in the analysis, one can multiply $X(t)$ by RTT to get the window size in terms of packets (or MSS).

[3] The linear growth is known to hold for TCP connections where the round-trip time is almost constant, or varies independently of the window size. The growth of $X(t)$ stops being linear when the RTT is correlated to the window size, see [5], [8]

[4] Usually ν is equal to one half, but we consider a more general scenario to account for other possible Additive Increase Multiplicative Decrease (AIMD) flow control mechanisms.

The model we propose in this paper is a fluid model that studies an AIMD flow control mechanism. It is then an extension of AIMD fluid models in the literature to scenarios where times between loss events are generally distributed, not only constant and exponential. This extension allows us to generalize the well know *square root formula* for TCP throughput, and to prove that it still holds in case of general stationary ergodic loss processes. We also explain how to adapt a fluid model to the TCP protocol, in particular how to account for the discrete nature of TCP, the receiver window limitation, and the Timeout mechanism. To consider these latter TCP mechanisms, we use techniques similar to those introduced in [35]. As a consequence, our model can be seen as an extension of [35] to scenarios where times between loss events are generally distributed not only constant. If we consider constant times between losses, we must obtain very close, if not the same, throughput as that obtained by [35]. Our experimental results validate this claim for loss rates ranging from few losses per 10,000 packets to few losses per 100 packets. As for higher loss rates, we expect our model to inherit the same performance limitation as [35] since finally both works deploy the same modeling for the Timeout mechanism. Modeling TCP performance under very high loss rates is not the main objective of this paper.

II. The main results

To compute the throughput, we use the following expression for the stationary regime of the process defined by (1):

$$X_n^* = \alpha \sum_{k=0}^{\infty} \nu^k S_{n-1-k}. \tag{2}$$

Next, we present various types of conditions under which (2) describes the unique stationary regime of our system.

A. The stationary regime

We consider the dynamic equation (1) under either one of the following assumptions:

- Θ_1: The process S_n is stationary-ergodic with $0 < \mathbb{E}[S_0] < \infty$. The distribution of the process S_n does not depend on X_n. We may thus construct on the same probability space a family of processes $X_n(x)$, $x \geq 0$ indexed by the initial state $X_0 = x$, such that all have the same inter-loss times S_n.
- Θ_2: The process (S_n, X_n) is stationary-ergodic with $0 < \mathbb{E}[S_0] < \infty$. Moreover, there is a unique stationary-ergodic regime that solves the dynamic equation (1).
- Θ_3: There is a stationary ergodic sequence η_n such that S_n can be represented as
$$S_n = S(X_n, \eta_n).$$
Equation (1) then becomes the so-called "stochastic recursive equation" (see [11]) of the form $X_{n+1} = f(X_n, \eta_n)$, where $f(X_n, \eta_n) = \nu X_n + \alpha S(X_n, \eta_n)$. We assume that f is nondecreasing in X_n and that S is nonincreasing in its first argument. Moreover, $\mathbb{E}[S(0, \eta_0)] < \infty$, and $0 < \mathbb{E}[S(a, \eta_0)]$, for some constant a.

Remark 1: An example in which even in presence of complex loss processes, Assumption Θ_1 holds, is given in [40]. A TCP-friendly application is considered in which the transmission rate of packets is constant, but the variations of the throughput are implemented by varying the packet size. This makes the TCP throughput independent of the distribution of the process of losses of packets.

Remark 2: The monotonicity of S in condition Θ_3 is quite natural. It reflects the fact that the time till the next loss tends in general to decrease as the window size increases, since there are more packets in the network and thus there are more chances for losses.[5] The condition $\mathbb{E}[S(0, \eta_0)] < \infty$, together with the monotonicity of S in the first argument, implies that $\mathbb{E}[S(x, \eta_0)]$ is finite for all x. The condition that $0 < \mathbb{E}[S(a, \eta_0)]$ for some constant a guarantees that there cannot be clusters of infinitely many simultaneous losses.

Proposition 1: Under either one of the assumptions Θ_1, Θ_2 or Θ_3, there is a unique stationary ergodic regime given by expression (2). Moreover, under Θ_1 or Θ_3, if the transmission rate evolution starts from an arbitrary rate X_0, it will converge almost surely to the above stationary regime,

$$\lim_{n \to \infty} |X_n - X_n^*| = 0, \quad P - \text{a.s.} \tag{3}$$

Proof: Under assumption Θ_1, equation (1) is a particular case of stochastic linear difference equations [12], [18]. Since the sequence of inter-loss times is stationary ergodic, it follows from Theorem 2A in [18] (and assuming that $0 < \nu < 1$ and that $0 < \mathbb{E}[S_n] < \infty$; see the Appendix of [2] for more details) that equation (1) has a stationary solution given by (2). Moreover, (3) follows from the results in [12], [18].

Under Assumption Θ_2, we see that X_n^* as defined in (2), is stationary ergodic, since it is a function of a stationary ergodic sequence. Moreover, the sum is well defined (since all summands are nonnegative), and it has finite expectation. Therefore, it is almost surely finite. Since under Θ_2 there is a unique stationary regime for (1), it has to be given by X_n^*.

Finally we consider Θ_3. We use a Loyns-type scheme [25] to show that X_n^* as defined in (2) is stationary and ergodic. We then show that it is finite and that X_n converges to the stationary regime from any initial state. Define the process $X_n^{(k)}$ to be a solution of (1) obtained with the initial condition $X_{-k}^{(k)} = 0$. Then it follows from the monotonicity of f in the first argument that for each fixed n, $X_n^{(k)}$ is monotone nondecreasing in k (for each sample, and for all $n > -k$). Note also that

$$X_n^{(k)} = \alpha \sum_{m=0}^{n+k-1} \nu^m S_{n-m-1}.$$

Thus the limit as $k \to \infty$ of $X_n^{(k)}$ equals X_n^*, and it is stationary ergodic (since it is a function of the stationary ergodic sequence η).

Due to the monotonicity of S, the sequence $X_n^{(k)}$ with a fixed initial state $X_0^{(k)} = x$, is bounded by the sequence $\hat{X}_n^{(k)}$ obtained by

$$\hat{X}_{n+1}^{(k)} = \nu \hat{X}_n^{(k)} + \alpha S_n(0, \eta_n)$$

with the same initial state $\hat{X}_0^{(k)} = x$. As the latter sequence is finite almost surely (assumption Θ_1 holds for that sequence),

[5] One exception could be the case of wide area networks where lot of traffic is multiplexed and where the loss process seen by a TCP connection can be approximated by a homogenous Poisson process independent of the window size, see the long-distance connection in the experimentation part.

it follows that $X_n^{(k)}$ is also almost surely finite. Moreover, by taking the limit as k tends to infinity, we also see that X_n^* is bounded by a stationary process which is finite almost surely, and therefore the process X_n^* is also finite almost surely.

Next we establish (3) under assumption Θ_3. Let X_n correspond to the process with initial state $X_0 = x$, and let X'_n correspond to the process with initial state $X'_0 = x'$. Without loss of generality, assume that $x' > x$. Then by the fact that f is monotone nondecreasing in the first argument, $X'_n \geq X_n$ for all n. On the other hand, from the fact that S is non increasing in its first argument, we have $X'_{n+1} - X_{n+1} = \nu(X'_n - X_n) + \alpha(S(X'_n, \eta_n) - S(X_n, \eta_n)) \leq \nu(X'_n - X_n)$. We conclude that $|X'_{n+1} - X_{n+1}| \leq \nu(X'_n - X_n)$, from which (3) follows. This implies the uniqueness of the stationary regime. □

Remark 3: Note that assumption Θ_1 implies assumption Θ_3 which in turns implies assumption Θ_2. Thus, Θ_2 is the weakest assumption under which we obtain the expression (2) for the stationary regime, whereas Θ_3 is the weakest assumption under which we obtain the convergence to the stationary regime (3).

Throughout the rest of the paper, we shall only consider the stationary ergodic regime, and our results will thus hold under assumption Θ_2.

B. The computation of the first two moments of X_n^* and the throughput of TCP

Here we compute the expectation and the second moment of the TCP transmission rate at the instants of losses as well as the TCP throughput.

Proposition 2: Let $\lambda = 1/\mathbb{E}[S_n]$ be the intensity of the loss process and let $R(k) = \mathbb{E}[S_n S_{n+k}]$ be the correlation function of the process $\{S_n\}_{n=-\infty}^{+\infty}$. Then,

$$\mathbb{E}[X_n^*] = \frac{\alpha}{\lambda(1-\nu)}, \tag{4}$$

$$\mathbb{E}\left[(X_n^*)^2\right] = \frac{\alpha^2}{1-\nu^2}[R(0) + 2\sum_{k=1}^{\infty} \nu^k R(k)]. \tag{5}$$

Remark 4: We note the remarkable insensitivity property, that $\mathbb{E}[X_n^*]$ does not depend on the correlation between inter-loss times nor on their moments of order greater than one.

Proof: To compute (4) and (5), we use the expression (2) for the stationary regime.

$$\mathbb{E}[X_n^*] = \alpha \sum_{k=0}^{\infty} \nu^k \mathbb{E}[S_{n-1-k}] = \frac{\alpha}{\lambda} \sum_{k=0}^{\infty} \nu^k = \frac{\alpha}{\lambda(1-\nu)}$$

Similarly, we obtain

$$\begin{aligned}
\mathbb{E}\left[(X_n^*)^2\right] &= \mathbb{E}\left[\alpha \sum_{j=0}^{\infty} \nu^j S_{n-1-j} \alpha \sum_{k=0}^{\infty} \nu^k S_{n-1-k}\right] \\
&= \alpha^2 \mathbb{E}\left[\sum_{k=0}^{\infty} \sum_{j=0}^{k} \nu^j S_{n-1-j} \nu^{k-j} S_{n-1-k+j}\right] \\
&= \alpha^2 \sum_{k=0}^{\infty} \sum_{j=0}^{k} \nu^k \mathbb{E}[S_{n-1-j} S_{n-1-k+j}] \\
&= \alpha^2 \sum_{k=0}^{\infty} \nu^k \begin{cases} R(0) + 2\sum_{j=1}^{r} R(2j), & k = 2r, \\ 2\sum_{j=1}^{r} R(2j-1), & k = 2r-1. \end{cases}
\end{aligned}$$

By regrouping the terms of the last series, we get (5). □

Remark 5: The expectation computed in (4) is taken with respect to loss instants. This expectation is also referred to as Palm expectation in the context of point processes [7].

Next, by using (2) and the concept of the Palm probability, we proceed for the computation of the TCP throughput:

$$\overline{X} = \lim_{T \to \infty} \frac{1}{T} \int_0^T X(t) \mathrm{d}t.$$

Our main result is the following closed-form expression for the throughput as a function of the correlation function R, of the loss intensity λ, the linear increase factor α and the multiplicative decrease factor ν.

Proposition 3: The throughput of TCP is given by

$$\overline{X} = \lambda\alpha[\frac{1}{2}R(0) + \sum_{k=1}^{\infty} \nu^k R(k)]. \tag{6}$$

Proof: Since the process $X(t)$ is ergodic, the throughput is equal to the expectation of the transmission rate $\mathbb{E}[X(t)]$ at an arbitrary time point. To compute $\mathbb{E}[X(t)]$ one can use the following inversion formula (see e.g., [7] Ch.1 Sec.4)

$$\mathbb{E}[X(t)] = \lambda \mathbb{E}^0\left[\int_0^{T_1} X(\tau) \mathrm{d}\tau\right], \tag{7}$$

where $\mathbb{E}^0[\cdot]$ is an expectation associated with Palm distribution. In particular, $\mathbb{P}^0\{T_0 = 0\} = 1$. Now using formula (7) and expression (2), we can write

$$\begin{aligned}
\mathbb{E}[X(t)] &= \lambda \mathbb{E}^0\left[\int_0^{T_1} (\nu X_0 + \alpha\tau) \mathrm{d}\tau\right] = \lambda \mathbb{E}^0\left[\nu X_0 S_0 + \frac{\alpha}{2} S_0^2\right] \\
&= \lambda \mathbb{E}^0\left[\alpha\nu \sum_{k=0}^{\infty} \nu^k S_{-1-k} S_0\right] + \lambda\alpha 2 \mathbb{E}^0\left[S_0^2\right] \\
&= \lambda\alpha \sum_{k=0}^{\infty} \nu^{k+1} R(k+1) + \frac{\lambda\alpha}{2} R(0) \\
&= \lambda\alpha[\frac{1}{2}R(0) + \sum_{k=1}^{\infty} \nu^k R(k)]
\end{aligned}$$

□

Remark 6: Often the covariance function $C(k) = R(k) - \mathbb{E}[S_n]^2$ is used instead of the correlation function $R(k)$. Then, the formulas (5) and (6) become

$$\mathbb{E}\left[(X_n^*)^2\right] = \frac{\alpha^2}{1-\nu^2}[C(0) + 2\sum_{k=1}^{\infty} \nu^k C(k)] + \frac{\alpha^2}{\lambda^2(1-\nu)^2},$$

$$\overline{X} = \lambda\alpha[\frac{1}{2}C(0) + \sum_{k=1}^{\infty} \nu^k C(k)] + \frac{\alpha(1+\nu)}{2\lambda(1-\nu)}.$$

Define $A(t)$ as the number of packets transmitted on the TCP connection until time t, and $L(t)$ be the number of loss events until time t. The probability p of losing a packet is then simply

$$p = \lim_{t \to \infty} \frac{L(t)}{A(t)} = \lim_{t \to \infty} \frac{\lambda t}{\int_0^t X(\tau) \mathrm{d}\tau} = \frac{\lambda}{\overline{X}}. \tag{8}$$

This allows us to write our main result (6) in another form so as to grasp the influence of p and RTT on the throughput for general distribution of inter-loss times. Define the normalized correlation function: $\hat{R}(n) = \lambda^2 R(n)$. Then using (8) and $\alpha = 1/(bRTT^2)$, we get

$$\overline{X} = \frac{1}{RTT\sqrt{pb}} \sqrt{\frac{1}{2}\hat{R}(0) + \sum_{k=1}^{\infty} \nu^k \hat{R}(k)}.$$

If we define, similarly, the normalized covariance as $\hat{C}(k) = \lambda^2 C(k)$ (where $C(k)$ is defined in Remark 6) then we obtain the following formula for the TCP throughput:

$$\overline{X} = \frac{1}{RTT\sqrt{pb}}\sqrt{\frac{1+\nu}{2(1-\nu)} + \frac{1}{2}\hat{C}(0) + \sum_{k=1}^{\infty}\nu^k\hat{C}(k)}. \quad (9)$$

We conclude that for arbitrary stationary ergodic loss process, the throughput of TCP is inversely proportional to RTT and to the square root of the packet loss probability p. This constitutes the main finding of our model, where the classical square root formula is generalized to the case of stationary ergodic losses.

Remark 7: Note that (6) can also be rewritten in terms of the second moment of the transmission rate at loss instants,

$$\overline{X} = \frac{\lambda(1-\nu^2)}{2\alpha}\mathbb{E}\left[(X_n^*)^2\right].$$

From this expression we can conclude that constant inter-loss times lead to the smallest TCP throughput over all the set of stationary loss processes having the same intensity.

C. Examples of loss process

Now let us consider some important particular cases of the general loss process.

1) IID random losses (General Renewal Process): We model the loss process as a general renewal process. Namely, we assume that $S_n, n = ..., -1, 0, 1, ...$ are i.i.d. random variables. The formulas (4), (5) and (6) take the following form.

Proposition 4: Let $\{S_n\}_{n=-\infty}^{+\infty}$ be i.i.d. with $d := \mathbb{E}[S_n]$ and $d^{(2)} := \mathbb{E}\left[S_n^2\right]$. Then,

$$\mathbb{E}[X_n^*] = \frac{\alpha d}{1-\nu}, \quad (10)$$

$$\mathbb{E}\left[(X_n^*)^2\right] = \frac{\alpha^2}{1-\nu^2}\left[d^{(2)} + \frac{2\nu d^2}{1-\nu}\right], \quad (11)$$

$$\overline{X} = \mathbb{E}[X(t)] = \frac{\alpha}{d}\left[\frac{1}{2}d^{(2)} + \frac{\nu d^2}{1-\nu}\right]. \quad (12)$$

In particular, if the inter-loss times are exponentially distributed, we have

$$\overline{X} = \frac{\alpha d}{1-\nu}. \quad (13)$$

For $\nu = 0.5$, this is similar to the expression obtained in [30]. If the inter-loss times are deterministic, we get

$$\overline{X} = \frac{1+\nu}{2(1-\nu)}\alpha d \quad (14)$$

With a change of variables, the above expression is equivalent to the classical square root formula obtained in the literature for deterministic losses [24], [28], [35]:

$$\overline{X} = \frac{1}{RTT}\sqrt{\frac{3}{2bp}}, \quad (15)$$

where p is the probability that a TCP packet is lost. Indeed, substituting d in (14) by its value in (8), setting $\nu = 0.5$, and recalling that α is equal to $1/(bRTT^2)$, we get the square root formula in (15). This can also be obtained from (9), as $\hat{C}(k) = 0$ for all k in the case of deterministic inter-loss times.

Remark 8: We note from (12) that the throughput of TCP can be expressed as a constant that only depends on the average time between loss events, plus a term that grows *linearly* with the variance of inter-loss times. Hence, the more variable the times between losses, the higher the throughput. When the loss events are highly bursty (which implies a large variance of inter-loss times), assuming that the loss process is Poisson [30] or deterministic yields a non-negligible underestimation of TCP throughput. Similarly, assuming that the loss process is Poisson when it is close to deterministic leads to an overestimation of TCP throughput.

2) Correlated losses modeled as a Markovian Arrival Process: In this section we consider correlated losses which are modeled by Markovian Arrival Process (MAP) [26], [32], [33]. It was shown in [6] that for a given general point process, there is a sequence of MAPs which converges to the point process in distribution. In particular, this implies that in principle the general point process can be approximated by appropriate MAPs. Furthermore, the PH-renewal process [33] and the Markov Modulated Poisson Process (MMPP) [14] are particular cases of the Markovian arrival process.

Let us briefly review the definition and some properties of the Markovian Arrival Process. Let $N(t)$ be a counting process associated with MAP, that is, $N(t)$ is the number of arrivals (or losses in our setting) in the interval $(0, t]$. Also let $J(t)$ be an auxiliary state variable. Then MAP can be described in terms of a two-dimensional Markov process $\{N(t), J(t)\}$ on the state space $\{(i, j)|i \geq 0, 1 \leq j \leq m\}$ with the following infinitesimal generator

$$Q = \begin{bmatrix} C & D & 0 & 0 & \cdots \\ 0 & C & D & 0 & \cdots \\ 0 & 0 & C & D & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where the matrix $C \in \mathbb{R}^{m\times m}$ governs the transition of the process J without arrival (loss) and it has negative diagonal elements and nonnegative off-diagonal elements. The matrix $D \in \mathbb{R}^{m\times m}$ governs the transitions of J with the simultaneous arrivals and it has nonnegative elements. Thus, the underlying Markov process $J(t)$ has the following infinitesimal generator $\overline{Q} = C + D$. Further, we assume that $\overline{Q} \neq C$ and that C is a stable matrix. This ensures that $J(t)$ does not get absorbed in a class of states in which arrivals stop. When $J(t) = i$, the rate of transitions to state $j \neq i$ is $\overline{Q}_{ij}$. If such a transition occurs then an arrival occurs simultaneously with the transition with probability $D_{ij}/(-C_{ii} - D_{ii})$. Note that MAP becomes MMPP with infinitesimal generator R and arrival rate matrix Λ, if we take $C = R - \Lambda$ and $D = \Lambda$.

Let $\{S_n\}_{n=1}^{\infty}$ be the sequence of inter-arrival times for MAP, and let $\{J_n\}_{n=1}^{\infty}$ be the sequence of states of the underlying Markov process at the arrival epochs. Then $\{J_n, S_n\}_{n=1}^{\infty}$ is a Markov renewal process [19] with the following transition probability matrix [33]:

$$F(x) = \left(\int_0^x \exp\{Cu\}\mathrm{d}u\right)D = (I - \exp\{Cx\})(-C)^{-1}D.$$

Note that $T = F(\infty) = -C^{-1}D$ is a transition probability matrix of a discrete time Markov chain embedded at the instants of arrivals. Let μ be its stationary distribution. If we take the initial distribution of the underlying Markov chain $J(t)$ as μ, the arrival process becomes event-stationary. The event-stationary version of MAP has the following joint distribution

function for the inter-arrival times [22]:

$$F_{S_0\cdots S_n}(x_0,...,x_n) = \mu \prod_{i=0}^{n} \{(I - \exp\{Cx_i\})T\}e. \quad (16)$$

Consequently, the joint Laplace-Stieltjes transform is given by

$$f(z_0,...,z_n) = \mathbb{E}\left[\exp\{-\sum_{k=0}^{n} z_k S_k\}\right] = \mu \prod_{k=0}^{n} \{(z_k I - C)^{-1} D\}e. \quad (17)$$

Using this Laplace-Stieltjes transform, we can easily compute the first two moments and the correlation function of the inter-arrival time process. Namely,

$$\mathbb{E}[S_n] = -\frac{d}{ds}(\mu(zI - C)^{-1}De)|_{z=0} = -\mu C^{-1}e, \quad (18)$$

$$\mathbb{E}\left[S_n^2\right] = \frac{d^2}{ds^2}(\mu(zI - C)^{-1}De)|_{z=0} = 2\mu C^{-2}e \quad (19)$$

$$\begin{aligned} R(k) &= \mathbb{E}[S_n S_{n+k}] = \frac{\partial^2}{\partial z_0 \partial z_k} f(z_0,...,z_k)|_{z_i=0} \\ &= \mu C^{-2} D T^{k-1} C^{-2} De. \end{aligned} \quad (20)$$

To derive the expression for the correlation function $R(k)$, we have used the following formula for the differentiation of an inverse matrix-valued function: $(A^{-1}(z))' = -A^{-1}(z)A'(z)A^{-1}(z)$ [10].

Now, we can calculate the first two moments of the process $\{X_n\}$ and the TCP throughput for a MAP loss process.

Proposition 5: Let the loss process $\{S_n\}$ be represented by MAP. Then,

$$\mathbb{E}[X_n^*] = -\frac{\alpha}{1-\nu}\mu C^{-1}e \quad (21)$$

$$\mathbb{E}\left[(X_n^*)^2\right] = \frac{\alpha^2}{1-\nu^2} 2\mu(C^{-2} + \nu C^{-2}D[I - \nu T]^{-1}C^{-2}D)e \quad (22)$$

Proof: The above formulas are immediately obtained from (4), (5) with the help of (18), (19), (20) and the following derivation

$$\sum_{k=1}^{\infty} \nu^k R(k) = \mu C^{-2} D \sum_{k=1}^{\infty} \nu^k T^{k-1} C^{-2} De$$

$$= \mu C^{-2} D \nu \sum_{k=0}^{\infty} \nu^k T^k C^{-2} De = \nu \mu C^{-2} D[I - \nu T]^{-1} C^{-2} De$$

□

Proposition 6: Let $\{S_n\}$ be a Markovian Arrival Process. Then, the throughput of TCP is given by

$$\overline{X} = -\frac{\alpha}{\mu C^{-1} e}\mu(C^{-2} + \frac{1}{2}C^{-2}D[I - \nu T]^{-1}C^{-2}D)e. \quad (23)$$

D. Higher moments of the transmission rate

Similarly to the way with which we compute the throughput of TCP, one can find the expression of any moment of the TCP transmission rate in the stationary regime. Of particular interest is the second moment which tells us how much the transmission rate of TCP oscillates. This is useful for the design of TCP-friendly transport protocols for multimedia applications. Multimedia applications are known to require small oscillations in the transmission rate [16], [17], [40], while the TCP-friendly requirement urges them to transmit their packets in a way that their average rate is no more than the average rate of a TCP connection.

Using the Palm inversion formula as in the proof of Proposition 3, the moment of order k of $X(t)$ in the stationary regime can be written as follows,

$$\mathbb{E}\left[X^k(t)\right] = \lambda \mathbb{E}^0\left[\int_0^{T_1} (\nu X_0 + \alpha\tau)^k \mathrm{d}\tau\right].$$

Developing the term inside the integral using the Binomial formula then integrating, we get the following expression for the k-th moment of the transmission rate of TCP,

$$\mathbb{E}\left[X^k(t)\right] = \lambda \sum_{i=0}^{k} \frac{C_k^i \nu^i \alpha^{k-i}}{k-i+1} \mathbb{E}^0\left[X_0^i S_0^{k-i+1}\right]. \quad (24)$$

We still have to compute the expectations $\mathbb{E}^0\left[X_0^i S_0^{k-i+1}\right]$ for $i = 0$ to k. These expectations can be easily computed by using the expression of X_0 in the stationary regime given in (2). We show next the expressions of these expectations for $k = 2$.

The second moment of $X(t)$, and hence the variance, requires the expressions of $\mathbb{E}\left[S_0^3\right]$, $\mathbb{E}\left[X_0 S_0^2\right]$ and $\mathbb{E}\left[X_0^2 S_0\right]$. Using (24), this second moment is equal to

$$\mathbb{E}\left[X^2(t)\right] = \lambda\left(\nu^2 \mathbb{E}\left[X_0^2 S_0\right] + \nu\alpha \mathbb{E}\left[X_0 S_0^2\right] + \frac{\alpha^2}{3}\mathbb{E}\left[S_0^3\right]\right). \quad (25)$$

$\mathbb{E}\left[S_0^3\right]$ is a characteristic of the loss process. Using the expression of X_0 in (2), the other two expectations can be expressed as a function of the auto-correlation functions of the loss process. We have,

$$\begin{aligned} \mathbb{E}\left[X_0 S_0^2\right] &= \alpha \sum_{k=0}^{\infty} \nu^k \mathbb{E}\left[S_0^2 S_{-1-k}\right] \\ \mathbb{E}\left[X_0^2 S_0\right] &= \alpha^2 \mathbb{E}\left[\sum_{k=0}^{\infty} \nu^k S_{-1-k} \sum_{j=0}^{\infty} \nu^j S_{-1-j} S_0\right] \end{aligned}$$

$$= \alpha^2 \sum_{k=0}^{\infty} \nu^{2k}\left(\mathbb{E}\left[S_0 S_{-1-k}^2\right] + 2\sum_{j=1}^{\infty} \nu^j \mathbb{E}\left[S_0 S_{-1-k} S_{-1-k-j}\right]\right).$$

E. Bounds for the model with transmission rate limitation

In the previous sections we did not include in the modeling the fact that TCP transmission rate may stop growing when the congestion window exceeds the window advertised by the receiver. This latter quantity corresponds to the maximum number of packets that can wait at the destination before being handed to the application [30], [38]. We consider in this section that the transmission rate of TCP is limited by a maximal value M. We take $\nu = 0.5$ since TCP divides the minimum of the receiver window and the congestion window by two upon congestion. Note here that the transmission rate can be limited by other factors such as the buffer size at the source or the available bandwidth in the network. The difference from the case where the limitation is due to the receiver window is in the reduction factor which can be less than two. For example, when the limitation is caused by the available bandwidth, some packets of the TCP connection are stored in the buffer at the bottleneck router at the moment of congestion, so dividing the congestion window by two will free some of these packets, will reduce the round-trip time of the connection, but the ratio of the window size divided and the round-trip time will stay larger than half the available bandwidth due to those packets of the connection backlogged in the bottleneck router.

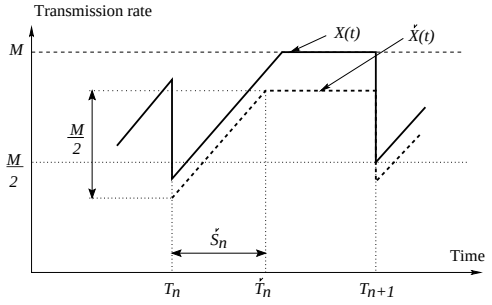

Fig. 1. TCP transmission rate evolution with limitation

The example of TCP transmission rate evolution we consider in this section is presented in Fig. 1. The stochastic difference equation (1) is respectively modified to the following form

$$X_{n+1} = M \wedge (\frac{1}{2}X_n + \alpha S_n), \tag{26}$$

where $\wedge$ stands for the *minimum* operation. The model becomes nonlinear and it is probably not possible to derive the explicit expressions of $\overline{X}$ for the general loss process.[6] We use here the results of the previous sections to obtain bounds on the performance measures instead.

Before deriving the bounds we shall state a stability result for the process X_n.

Theorem 1: Assume that $\{S_n\}$ is a stationary process. Then there exists a stationary process $\{X_n^*\}$ defined on the same probability space, and satisfying the recursion (26). Furthermore, for any initial state X_0, we have P-a.s.

$$\lim_{n\to\infty} |X_n - X_n^*| = 0.$$

Proof: The proof can be easily made using a Loynes-type construction similar to that used for Proposition 1. We refer to [3] for details. □

Now we note that equation (26) can be rewritten as

$$X_{n+1} = \frac{1}{2}X_n + \alpha S_n \wedge (M - \frac{1}{2}X_n).$$

Since $0 \le X_n \le M$, we have

$$\frac{1}{2}X_n + \alpha S_n \wedge \frac{M}{2} \le X_{n+1} \le \frac{1}{2}X_n + \alpha S_n \wedge M.$$

The above estimates prompt us to derive lower and upper bounds on the throughput, using the next auxiliary stochastic processes defined on the same probability space as X_n:

$$\check{X}_{n+1} = \frac{1}{2}\check{X}_n + \frac{M}{2} \wedge (\alpha S_n) = \frac{1}{2}\check{X}_n + \alpha(\frac{M}{2\alpha} \wedge S_n), \tag{27}$$

and

$$\hat{X}_{n+1} = \frac{1}{2}\hat{X}_n + M \wedge (\alpha S_n) = \frac{1}{2}\hat{X}_n + \alpha(\frac{M}{\alpha} \wedge S_n). \tag{28}$$

Proposition 7: Let $\{S_n\}$ be a stationary stochastic point process. Assume that $X_0 = \check{X}_0 = \hat{X}_0$. Then for all $n \ge 0$, $\check{X}_n \le X_n \le \hat{X}_n$. Moreover, $2\alpha\check{d} \le \mathbb{E}[X_n] \le 2\alpha\hat{d}$, where the expectation $\mathbb{E}[X_n]$ is taken with respect to the stationary regime, $\check{d} = \mathbb{E}\left[\check{S}_n\right]$, $\hat{d} = \mathbb{E}\left[\hat{S}_n\right]$, and where $\check{S}_n := \frac{M}{2\alpha} \wedge S_n$, $\hat{S}_n := \frac{M}{\alpha} \wedge S_n$.

Proof: We show by induction that $\check{X}_n \le X_n$. It holds for $n = 0$. Assume it holds for $n = k$. Then, consider two cases $S_k \le \frac{M}{2\alpha}$ and $S_k > \frac{M}{2\alpha}$. For $S_k \le \frac{M}{2\alpha}$, one has

$$X_{k+1} = \frac{1}{2}X_k + \alpha S_k \ge \frac{1}{2}\check{X}_k + \alpha S_k = \check{X}_{k+1}.$$

And if $S_k > \frac{M}{2\alpha}$, then

$$\begin{aligned} X_{k+1} &= \frac{1}{2}X_k + (M - \frac{1}{2}X_k) \wedge (\alpha S_k) \\ &\ge \frac{1}{2}X_k + (M - \frac{1}{2}X_k) \wedge (\frac{M}{2}) \\ &= \frac{1}{2}X_k + \frac{M}{2} \ge \frac{1}{2}\check{X}_k + \frac{M}{2} = \check{X}_{k+1} \end{aligned}$$

The first inequality is true, since $X_k \le M$. Hence, $\check{X}_{k+1} \le X_{k+1}$ and according to the induction principle, the inequality $\check{X}_n \le X_n$ holds for all $n \ge 0$. Consequently, $\mathbb{E}\left[\check{X}_n\right] \le \mathbb{E}[X_n]$ for all $n \ge 0$.

Since by the results of [18] and Theorem 1 both processes $\{\check{X}_n\}$ and $\{X_n\}$ converge to the stationary regime, we can let n go to infinity. This results in the lower bound which holds for any initial state of the two processes.

The upper bound is obtained in a similar manner by using the auxiliary process (28).

□

Now we calculate the lower and upper bounds on the throughput.

Proposition 8: Let $\check{d}^{(2)} := \mathbb{E}\left[(\check{S}_n)^2\right]$, $\check{R}(k) := \mathbb{E}\left[\check{S}_{n-k}S_n\right]$ and $\hat{d}^{(2)} := \mathbb{E}\left[(\hat{S}_n)^2\right]$, $\hat{R}(k) := \mathbb{E}\left[\hat{S}_{n-k}S_n\right]$. Then, the lower and upper bounds on the throughput are given by

$$\overline{X} \ge \alpha\lambda\left(\check{R}(0) - \frac{1}{2}\check{d}^{(2)} + \sum_{k=0}^{\infty}\frac{1}{2^{k+1}}\check{R}(k+1)\right), \tag{29}$$

$$\overline{X} \le \alpha\lambda\left(\hat{R}(0) - \frac{1}{2}\hat{d}^{(2)} + \sum_{k=0}^{\infty}\frac{1}{2^{k+1}}\hat{R}(k+1)\right). \tag{30}$$

Proof: To obtain the lower bound on the throughput, we again use the auxiliary process (27). Suppose that $\{X_n\}$ is in the stationary regime and define

$$\check{X}(t) = \begin{cases} \frac{1}{2}\check{X}_n^* + \alpha t, & t \in [T_n, \check{T}_n], \\ \frac{1}{2}\check{X}_n^* + \alpha\check{S}_n, & t \in [\check{T}_n, T_{n+1}], \end{cases} \tag{31}$$

where $\check{T}_n = T_n + \check{S}_n$ (see Fig. 1). Similarly to (2), one can write the expression for the stationary version of $\{\check{X}_n\}$, that is $\check{X}_n^* = \alpha\sum_{k=0}^{\infty}(\frac{1}{2})^k\check{S}_{n-1-k}$. Using (31) and the above expression for $\check{X}_n^*$, we obtain the lower bound

$$\begin{aligned} \overline{X} &= \lambda\mathbb{E}^0\left[\int_0^{T_1} X(t)\mathrm{d}t\right] \ge \lambda\mathbb{E}^0\left[\int_0^{T_1}\check{X}(t)\mathrm{d}t\right] \\ &= \lambda\mathbb{E}^0\left[\int_0^{\check{S}_0}(\frac{\check{X}_0^*}{2} + \alpha t)\mathrm{d}t + \int_{\check{S}_0}^{S_0}(\frac{\check{X}_0^*}{2} + \alpha\check{S}_0)\mathrm{d}t\right] \\ &= \lambda\mathbb{E}^0\left[\frac{1}{2}\check{X}_0^*\check{S}_0 + \frac{\alpha}{2}\check{S}_0^2 + (\frac{1}{2}\check{X}_0^* + \alpha\check{S}_0)(S_0 - \check{S}_0)\right] \\ &= \lambda\mathbb{E}^0\left[\frac{1}{2}\check{X}_0^*S_0 + \alpha\check{S}_0S_0 - \frac{\alpha}{2}\check{S}_0^2\right] \\ &= \lambda\mathbb{E}^0\left[\frac{1}{2}\alpha\sum_{k=0}^{\infty}(\frac{1}{2})^k\check{S}_{-1-k}S_0 + \alpha\check{S}_0S_0 - \frac{\alpha}{2}\check{S}_0^2\right] \\ &= \alpha\lambda\left(\sum_{k=0}^{\infty}(\frac{1}{2})^{k+1}\check{R}(k+1) + \check{R}(0) - \frac{1}{2}\check{d}^{(2)}\right). \end{aligned}$$

Now, by using the auxiliary process (28), one can calculate the upper bound on the throughput in a similar way. □

Note that the two bounds given in Proposition 8 coincide with the throughput given by (6) as $M\lambda/\alpha \to \infty$ (due to large M or to high loss rate). However, when $M\lambda/\alpha \to 0$, the upper bound provided in (30) goes to $2M$. Therefore, we propose to

[6] We refer to [4] for an illustration on how much this derivation is difficult even in the simple case of a homogenous Poisson loss process.

take as an upper bound on the TCP throughput the minimum between M and the upper bound given in (30). As for the lower bound, it converges to the following expression

$$\overline{X} \simeq M - \frac{\lambda M^2}{8\alpha} \tag{32}$$

as $M\lambda/\alpha \to 0$. As was shown in [2], [35], the expression (32) appears to be a very good approximation of the throughput when the maximal rate M is frequently reached. It is the throughput obtained by TCP when the transmission rate reaches its maximum value M between each two losses.

Let us specify the two bounds on the throughput for the cases of Poisson and IID losses. We refer to [2] for the case of losses driven by a MAP process.

Assume first that the loss process is Poisson. Then formulas (29) and (30) give the following bounds on the throughput

$$\frac{2\alpha}{\lambda}\left(1 - e^{-M\lambda/2\alpha}\right) \leq \overline{X} \leq \frac{2\alpha}{\lambda}\left(1 - e^{-M\lambda/\alpha}\right).$$

Note that $2\alpha/\lambda$ is the TCP throughput in the case of Poisson losses and an infinite maximum window size M (see Subsection II-C.1).

Consider next the more general case of an IID loss process. The correlation functions $\check{R}(k)$ and $\hat{R}(k)$ are simply equal to $d\check{d}$ and $d\hat{d}$ respectively. We have then the following bounds,

$$\alpha\lambda\left(\check{R}(0) - \frac{1}{2}\check{d}^{(2)} + d\check{d}\right) \leq \overline{X} \leq \alpha\lambda\left(\hat{R}(0) - \frac{1}{2}\hat{d}^{(2)} + d\hat{d}\right)$$

F. Modeling conservative Timeouts

We were assuming till now that losses were quickly detected. This is indeed the case when losses are detected via duplicate ACKs (the Fast Retransmit algorithm [38]) or via a fine-granularity correctly-set retransmission timer. However, most TCP implementations use a coarse-granularity timer (500ms in unix implementations) for the detection of losses in the case when three duplicate ACKs are not received. This coarse-granularity together with the back-off mechanism of the retransmission timer in case of retransmission losses introduce some idle times during which the congestion window of TCP is not increasing and the transmission rate is approximately equal to zero. We call these losses followed by an idle time before the resumption of the transmission *Timeout losses* (TO). The idle time separates the loss of a packet and the receipt of the ACK for its retransmission. This includes any back-off of the retransmission timer due to retransmission loss. Losses which are detected quickly without the need for an idle period are called TD losses (TD for three duplicate ACKs). As shown in our experimentations, TO losses can be quite frequent. For instance, we refer to Tables I, II and III, where we show statistics on three long-lived TCP connections. The 10th columns in these tables (labelled Q) indicate the percentage of losses which are of TO type. This number is sometime non-negligible for different reasons: the high loss rate that a TCP connection may encounter at some hours during the day (the column labelled p), the fact that multiple packets can be lost upon a congestion event, and finally the inaccuracy in the estimation of the TCP retransmission timer. The same finding has been reported in [35]. We explain in this section how one can include these idle times into our explicit expression for the throughput, even though we believe that this phenomenon will be of negligible importance in the future given the different enhancements proposed recently to enhance the TCP error recovery phase, e.g. SACK [27], Limited Transmit [1].

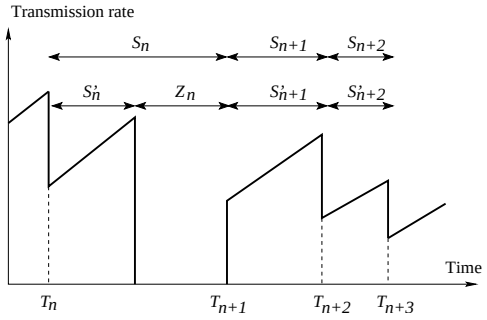

Fig. 2. A model for TCP with TO and TD losses

Let Z_n be the duration of the idle period after loss event n. Z_n is equal to 0 if the loss is of TD type, and is greater than zero if the loss is of TO type. Let $Z = \mathbb{E}[Z_n | Z_n > 0]$ denote the average duration of the idle periods after TO losses, and let $Q = \mathbb{P}\{Z_n > 0\}$ denote the probability that a loss is of type TO. Define T_n as the instant at which the transmission rate resumes its increase after the nth loss (TD or TO), and let $S_n = T_{n+1} - T_n$. If the nth loss is of TO type, the connection gets in an idle period at time $T_n - Z_n$ until time T_n where the transmission is resumed. During this idle period, the transmission rate is equal to zero (we are interested in the TCP throughput). We assume that after the idle period, the transmission rate jumps directly to half its value before the TO loss. This is justified by the fact that the slow start phase after Timeout is fast compared to the linear increase phase. We depict in Fig. 2 a sample of the transmission rate evolution in presence of TO losses according to our model.

Define now the sequence (see Fig. 2)

$$S'_n = \begin{cases} S_n & \text{if the loss is } TD \\ S_n - Z_n & \text{if the loss is } TO \end{cases},$$

and assume it to be stationary ergodic.[7] Let $\lambda' = 1/\mathbb{E}[S'_n]$ and let $\hat{C}'(k)$ denote the normalized covariance function of the sequence $\{S'_n\}$. Using (9), the throughput of TCP when excluding Timeout intervals is given by

$$\overline{X'} = \frac{1}{RTT\sqrt{pb}}\sqrt{\frac{1+\nu}{2(1-\nu)} + \frac{1}{2}\hat{C}'(0) + \sum_{k=1}^{\infty}\nu^k\hat{C}'(k)}.$$

When including Timeouts, the throughput of TCP can be shown to be equal to (see [8] for proof):

$$\overline{X} = \frac{\overline{X'}}{1 + \lambda' QZ} = \frac{\overline{X'}}{1 + p\overline{X'}QZ}.$$

The last equality follows from the fact that $\lambda' = p\overline{X'}$.

Two general functions appear in the above modeling: Q and Z. To compute these functions, one needs to model the loss process at the packet level, i.e. how many packets are lost upon a congestion event, then to model the way with which TCP handles these packet losses. This is a complex problem strongly

[7] Note that the distribution of the time between the nth and the $(n+1)$th loss may depend on the type of nth loss (TD or TO). This however does not prevent the process S' of being stationary.

dependent on the version of TCP.[8] Some effort has been made in [35] to model the reaction of TCP to the first packet loss upon a congestion event.[9] Q and Z have been computed as a function of p, the packet loss probability, then validated with real experimentations over the Internet. The expression of Z is somehow general and holds for all versions of TCP. The expression of Q suites those versions of TCP that have an intelligent Fast Recovery phase, and that only timeout when the first packet lost upon congestion cannot be recovered by Fast Retransmit.[10] These expressions have shown good performance in modeling TCP Timeouts. This good performance added to the fact that our main focus in this work is on the distribution of inter-arrival times more than on Timeouts, motivated us to use the expressions of Q and Z proposed in [35]. This also allows a fair comparison of our model with [35].

III. Model Validation

Our work is mainly motivated by the fact that the loss of TCP packets over the Internet may present a more complex structure than the simple processes considered in the literature. To support this motivation, we run real long-lived TCP connections between several Internet sites. For each connection, we measure the instants of losses as well as some other statistics as the average round-trip time and the total number of packets transmitted. We study then our model under different assumptions on the type of the loss process. After that, we evaluate how well linear rate increase models approximate real TCP performance. We also evaluate our expression for Timeouts. At the end, we introduce a method to account for the discreteness of TCP. With this method, our model under deterministic inter-loss time assumption gives very close results to the detailed discrete model in [35].

A. Experimentation testbed

The experimentation has two purposes. First, to examine the loss process and to check whether it can be approximated by simple models. Second, to validate the TCP model.

We ran at different days during January 2000, three long-lived TCP transfers to three different machines. Each transfer consists of a continuous flow of data during a whole day. The source machine (`clope.inria.fr`) is running the New Reno version of TCP and is located at INRIA - Sophia Antipolis in the south of France. The destination machines are located respectively at the ESSI school at 1 km from INRIA (`nessie.essi.fr, 4 hops`), at the ENST school in Paris (`solo.enst.fr, 10 hops`), and at the University of South Australia (`linus.levels.unisa.edu.au, 22 hops`). TCP Packets are of 1460 Bytes size (excluding TCP and IP headers). The machine in Australia advertises a window of 22 packets and those at ESSI and ENST advertise a window of 44 packets. All machines implement the Delay ACK mechanism, so $\alpha = 1/(2RTT^2)$. For the simplicity of the exposition, we denote the three TCP connections by SD (Short Distance), MD (Medium Distance) and LD (Long Distance), respectively. The experimentation testbed is depicted in Fig. 3.

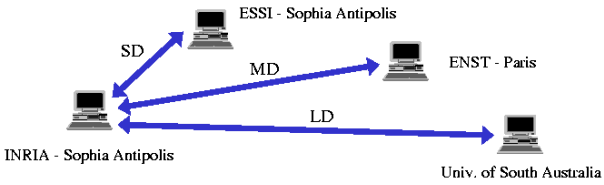

Fig. 3. The experimentation testbed

Data packets and the corresponding ACKs are captured with the `tcpdump` tool [36] at INRIA. Given the version of the TCP source, we developed a tool that looks at the trace of every connection and identifies the instants of window reductions (T_n). In the case of a loss detected via Timeout, the tool computes the duration of the Timeout period. In general, the developed tool determines the times S_n and S'_n, the packet loss probability p (number of loss events divided by the number of packets transmitted), the parameters Q and Z, the average RTT from the measured trace file, and the frequency with which the receiver window is reached ($\mathbb{P}\{X(t) = M\}$).

Each connection is run for multiple hours. Its total trace however is divided in short intervals of the order of minutes. This division is necessary since the loss process is certainly not stationary at the scale of the total trace duration. The stationarity of the loss process over time intervals of the order of minutes is judged a reasonable assumption according to [41]. We choose the intervals so that the number of loss events per interval is large enough for the characterization of the loss process to be accurate (around 500 loss events per interval). This gives us a set of trace files for each connection. For every trace file, we compute statistics on the TCP connection and on the loss process using our tool. We summarize the results in Tables I, II and III for respectively the SD, MD, and LD connections. Each row in the tables corresponds to a trace file. Columns present different information, which are from left to right: starting time of the trace in daytime hours (between 0 and 23), end time of the trace, number of bytes transmitted, number of packets transmitted, packet loss probability (p), the average RTT, the real throughput of the TCP connection in Kbps, the average rate of loss events (λ), the probability that a loss event results in Timeout (Q), the normalized covariance of times between loss events (when excluding Timeout intervals), and finally the probability that the receiver window is reached.[11]

For each trace file, we compare the real throughput of TCP to the one expected by our model as well as to the computed bounds. We also compare the real throughput to the one expected by [35]. We examine the validity of different models for the distribution of inter-loss times (deterministic, Poisson, iid, general correlated). The moments and correlation functions of inter-loss times are calculated from the trace files.

We study separately the effect of assumptions on the loss process and the correctness of our fluid model for TCP

[8] We refer to [13] for a discussion on how the different versions of TCP react to packet losses.

[9] TCP reacts either by timing out or by detecting the first loss by Fast Retransmit (three duplicate ACKs)

[10] This can be the case of the SACK version that does not support the Limited Transmit enhancement.

[11] Equivalently the fraction of time during which the TCP connection is transmitting at its maximum window.

Begin (hour)	End (hour)	Byte # $\times 10^3$	Packet #	Loss p (%)	RTT (ms)	Thrp (Kbps)	$\lambda = 1/\mathbb{E}[S_n]$ (1/s)	Q (%)	CoV (S'_n) (%)	$\mathbb{P}\{X(t) = M\}$ (%)
10.80	11.07	157125	115110	0.36	112	1280.90	0.4300	1.42	115.78	4.92
11.07	11.34	155500	113919	0.47	105	1293.96	0.5575	2.05	115.85	1.96
11.34	11.60	156030	114308	0.45	104	1314.47	0.5423	3.49	113.92	1.64
11.60	11.86	156963	114991	0.38	113	1346.41	0.4728	2.72	136.24	4.97
11.62	11.87	139052	101869	0.45	114	1254.14	0.5264	2.35	138.90	5.43
11.87	12.11	138739	101640	0.68	84	1271.68	0.7997	1.43	95.22	0.69
12.11	12.37	140466	102906	0.56	92	1226.88	0.6299	3.81	120.09	4.55
12.37	12.63	143625	105219	0.29	146	1234.12	0.3286	1.63	150.43	13.70
13.20	13.47	150529	110277	0.48	117	1205.46	0.5405	2.40	137.48	3.21
13.47	13.76	144508	105866	0.77	106	1101.02	0.7828	1.33	138.52	0.39
13.77	14.12	190662	139679	0.60	110	1187.22	0.6553	1.90	139.16	3.18

TABLE I

STATISTICS ON THE SHORT-DISTANCE CONNECTION

Begin (hour)	End (hour)	Byte # $\times 10^3$	Packet #	Loss p (%)	RTT (ms)	Thrp (Kbps)	$\lambda = 1/\mathbb{E}[S_n]$ (1/s)	Q (%)	CoV (S'_n) (%)	$\mathbb{P}\{X(t) = M\}$ (%)
14.52	15.43	219758	160995	1.40	141	532.37	0.6846	9.73	49.40	0
18.43	19.71	212720	155838	2.21	152	385.37	0.7817	20.42	55.29	0
19.71	20.32	227599	166739	0.67	125	834.90	0.5195	3.35	46.61	0
20.32	21.00	228578	167456	0.69	138	757.10	0.4823	9.44	76.64	1.38
21.00	21.52	229746	168312	0.51	119	1002.09	0.4710	1.38	40.69	0
21.52	22.08	228937	167719	0.57	123	915.53	0.4808	1.14	36.91	0
22.08	22.60	229427	168078	0.52	120	985.37	0.4735	2.26	38.53	0
22.60	23.11	229826	168371	0.52	114	1024.21	0.4957	1.57	37.82	0

TABLE II

STATISTICS ON THE MEDIUM-DISTANCE CONNECTION

Begin (hour)	End (hour)	Byte # $\times 10^3$	Packet #	Loss p (%)	RTT (ms)	Thrp (Kbps)	$\lambda = 1/\mathbb{E}[S_n]$ (1/s)	Q (%)	CoV (S'_n) (%)	$\mathbb{P}\{X(t) = M\}$ (%)
17.52	19.22	43816	32099	1.84	1245	57.19	0.0962	67.97	114.54	0.84
19.22	21.07	42369	31039	4.20	781	51.21	0.1970	66.33	80.94	0
21.07	22.29	44296	32451	2.70	662	80.37	0.1989	70.35	95.62	0.07
22.29	23.01	45036	32993	1.32	596	139.16	0.1687	59.49	105.83	2.53
23.01	0.34	43356	31762	2.92	684	72.83	0.1952	66.98	95.91	0.25
0.34	1.10	45005	32970	1.41	613	132.15	0.1717	61.75	123.76	3.94
1.10	1.55	45490	33326	0.59	647	220.88	0.1195	55.83	142.36	13.40
1.55	2.08	45121	33056	0.68	652	193.17	0.1214	51.54	140.92	12.56
2.08	2.45	45477	33316	0.43	578	269.68	0.1067	53.47	159.16	28.74
2.45	2.75	46035	33725	0.15	565	358.12	0.0515	39.62	138.10	54.79
2.75	3.05	45867	33602	0.23	579	332.29	0.0715	44.30	156.57	41.65
3.05	3.50	45506	33338	0.67	605	225.55	0.1387	61.16	199.21	25.28
3.50	3.80	46068	33750	0.19	564	356.79	0.0629	44.61	150.95	56.50
3.80	4.10	46065	33747	0.18	631	335.40	0.0555	52.45	134.85	46.88
4.10	4.41	45966	33675	0.21	584	331.94	0.0649	47.22	206.70	51.19
4.41	4.97	92305	67623	0.10	570	371.10	0.0366	38.35	176.68	66.19
4.97	5.23	46196	33843	0.12	564	378.07	0.0439	46.51	195.97	60.23
5.23	5.79	92480	67751	0.12	574	361.38	0.0429	40.90	158.22	57.88
5.79	6.10	46291	33913	0.14	567	357.64	0.0482	46.00	140.03	55.44
6.10	6.42	45655	33447	0.28	570	317.21	0.0825	41.05	142.73	32.28
6.42	6.79	45801	33554	0.39	564	292.21	0.1068	54.47	198.34	31.93
6.79	7.08	45887	33617	0.19	567	350.42	0.0610	46.87	161.51	55.99
7.09	8.08	185215	135688	0.06	659	401.16	0.0240	35.95	196.20	74.04
8.08	8.88	138616	101550	0.04	565	411.59	0.0155	38.09	189.07	78.25
8.88	9.31	45527	33353	0.42	689	238.65	0.0923	50.35	180.34	35.03
9.31	10.21	44724	32765	1.64	686	110.65	0.1667	68.46	123.66	1.93
10.21	12.09	44487	32591	2.91	976	52.66	0.1404	74.71	102.84	0
12.09	14.73	44377	32511	3.56	1169	37.35	0.1219	78.94	98.39	0
14.74	17.52	43433	31819	3.31	1355	35.41	0.1074	73.90	94.19	0

TABLE III

STATISTICS ON THE LONG-DISTANCE CONNECTION

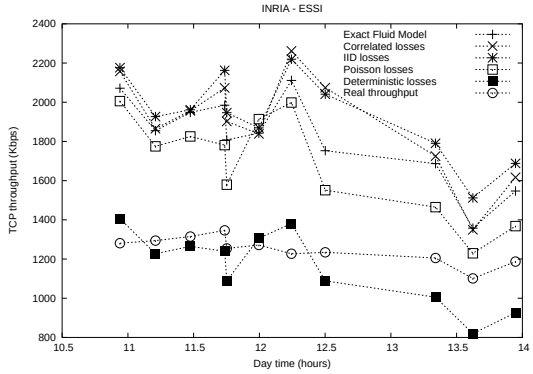

Fig. 4. Short-distance connection

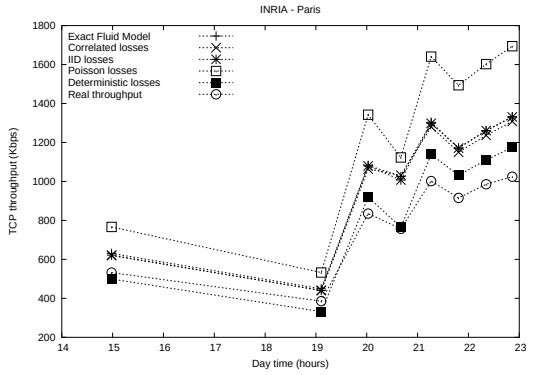

Fig. 5. Medium-distance connection

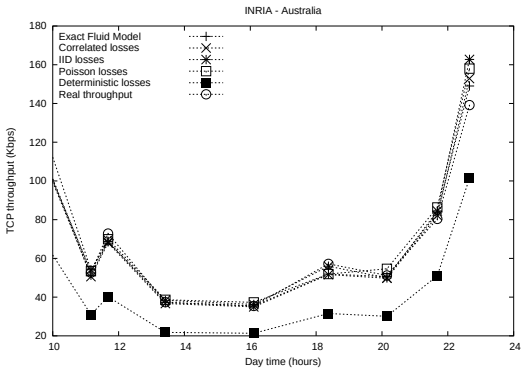

Fig. 6. Long-distance connection

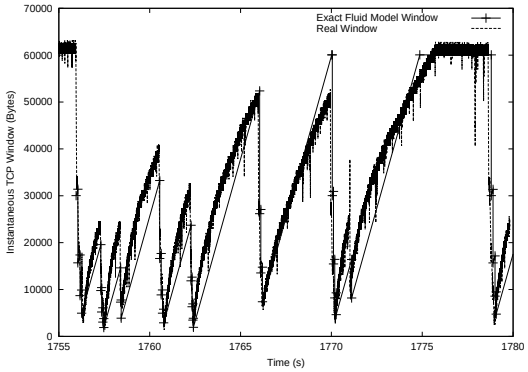

Fig. 7. Fluid model vs. real window on the short-distance connection

transmission rate evolution. The validation of the loss model was done as follows. We reconstruct for a given trace file the evolution in time of the proposed fluid model (i.e. the mechanism with linear increase and multiplicative decrease, silence time during Timeouts, and maximum limit on the congestion window; see Fig. 7). We call this process the *Exact Fluid Model* (EFM) and we compute exactly its throughput. This is done by computing the area below the transmission rate between two consecutive losses, then by summing all the areas and dividing the result by the transfer time. The EFM throughput is the throughput we are trying to estimate in our analysis and which we (and other authors that use linear rate increase fluid models to study TCP) are claiming that it represents the real TCP throughput. Our measure of how good a given model for the distribution of inter-loss times is, will be how close the throughput predicted by our closed-form expression (6) agrees with the EFM throughput.

Note that if the loss model is good according to the above criterion, we are still not guaranteed that the real throughput of TCP agrees with our throughput formulas. We expect the latter to be close to the real TCP throughput if the linear rate increase model is appropriate, which is not always the case as we will see later. In fact, on some paths as our LD connection, the increase of TCP rate is far from linear (sub-linear) which leads to a considerable throughput estimation error even if we use the right model for loss events in our general formula.

B. Validation of the model for losses

We measure how close is the throughput predicted by our closed-form expression (6) to the EFM throughput. Different loss processes are considered: deterministic, Poisson, general iid, general correlated. We plot the results in Fig. 4, 5 and 6 as a function of time for the three connections SD, MD and LD. For Timeouts, we measure the functions Q and Z directly from the traces rather than using those computed in [35]. The figures also show the real throughput of TCP. We clearly notice that our general model gives the same result as the exact fluid model although five terms are only considered in the infinite sum in formula (6). The iid model gives approximately the same result as the correlated model which means that losses are rarely correlated especially on the medium-distance and the long-distance connections. Some correlation can be seen on the short-distance connection, which is illustrated by the distance that separates the ”general correlated” throughput line from the ”iid” throughput line. Our analysis of the traces of the SD connection indicates indeed that loss events appear mostly in bursts. This burstiness of loss events can be seen in Fig. 7, where we plot the congestion window of the SD connection versus time during 25 seconds.

Consider now the Poisson and deterministic cases. The expression of the throughput in the iid case (12) states that at constant loss intensity, the throughput of TCP increases with the variance of inter-loss times. Thus, a comparison of the iid throughput line to the Poisson and deterministic throughput lines indicates how much times between loss events vary. On the SD connection, the variance of inter-loss times is more than that of the exponential distribution. This is caused by the bursty occurrence of losses we discovered on this connection as shown in Fig. 7. On such connection, one should expect that models assuming deterministic inter-loss times would give bad

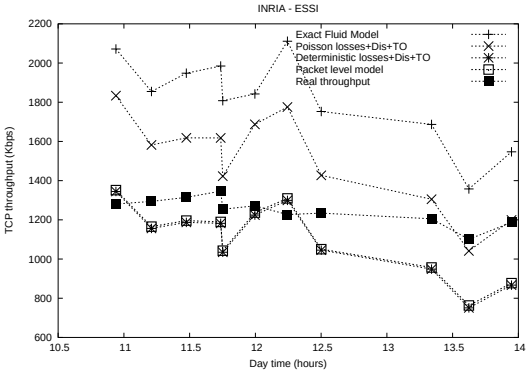

Fig. 8. Short-distance connection

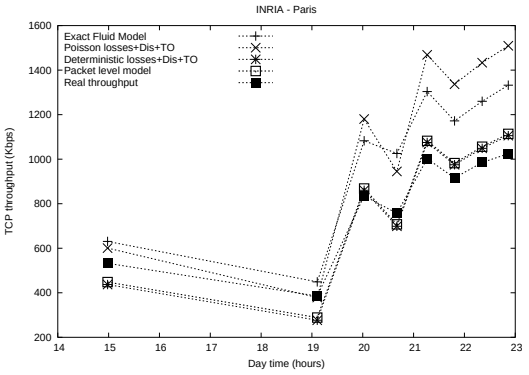

Fig. 9. Medium-distance connection

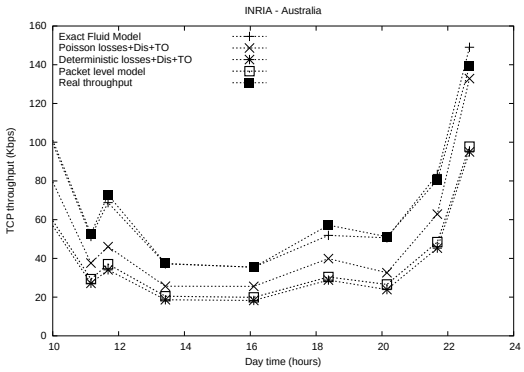

Fig. 10. Long-distance connection

results compared to the real throughput. Interestingly enough, it is not the case due to the problem of TCP rate sub-linearity that we will explain later. On the MD connection, we see that the variance of inter-loss times is closer to that of a deterministic distribution than that of an exponential distribution. Finally, on the LD connection, it is clear that losses occur according to a Poisson process, which is in some sense an expected result due to the high degree of multiplexing in Internet routers.

C. Validation of the model for TCP

We compare the exact fluid model (EFM) to real TCP on the three connections. Our objective is to test the validity of the linear growth assumption and the fluid assumption. The results are plotted in Fig. 8, 9, and 10. On the SD and MD connections, the exact fluid model overestimates real TCP and the overestimation increases with the real throughput. This is mainly due to the sub-linearity of TCP congestion window evolution, which can be seen in Fig. 7. TCP rate sub-linearity is due to the increase in the RTT with the congestion window, which in turn is due to the increase in the queuing time in bottleneck routers. We refer to [5] for an example on how this correlation between the RTT and the window size can be modeled. Unfortunately, as shown in [5], the modeling of the sub-linearity of TCP is complex even under the simple assumption that the loss process is Poisson. Note that this problem of sub-linearity does not exist on our LD connection where the window size is usually small, the propagation delay is large, and the TCP connection does not contribute considerably to the queuing time in network routers.

Another source of error in our model is the fluid approximation. In fact, the transmission rate of TCP does not increase continuously but rather jumps when the number of packets injected into the network increases by one. This is due to the Nagle algorithm [31] which prohibits a TCP source from injecting small packets into the network. However, the window size at the source can be assumed to change continuously with time between loss events. Fig. 11 explains the difference between the congestion window size (continuous line) and the number of packets in the network (dashed line). The expression of the throughput given by our fluid model corresponds to the average window size rather than to the number of packets in the network, so this expression has to be corrected to account for the area between the dashed and continuous lines.

We explain here how our model can be adapted to account for the discrete nature of TCP. Assume that $X(t)$ is measured in packets per unit of time. A good approximation is to shift down our process $X(t)$ by $1/(2RTT)$, then to subtract from the throughput the error caused by the number of packets lost upon congestion (Fig. 11). We introduce this last error since many packets can be lost upon a congestion event, and since we are interested in the computation of the throughput rather than the average sending rate. We follow [35] by assuming that on average, half of the window size is lost upon congestion. We also add the last RTT where some packets are transmitted until the detection of the congestion. On average, the number of these transmitted packets is equal to half the window size. We approximate the window size during these last two RTTs by the window size given by our fluid model upon loss events, i.e. $\mathbb{E}\left[X_n^*\right] * RTT$.[12] Thus, we subtract $\mathbb{E}\left[X_n^*\right] * RTT$ from the average integral of the transmission rate between two loss events. This gives the following corrected expression for the throughput:

$$\overline{X_d} = \overline{X} - \frac{1}{2RTT} - \frac{\mathbb{E}\left[X_n^*\right] * RTT}{\mathbb{E}\left[S_n\right]} = \overline{X} - \left(\frac{b}{2} + \frac{1}{1-\nu}\right) \alpha RTT$$

Using this correction, we compare in Fig. 8, 9, and 10 our model with deterministic inter-loss times (14) to the packet level model in [35]. We also plot in the figures the throughput obtained with our model under the assumption that losses are Poisson. We use for Q and Z the expressions computed in [35] rather than those obtained from measurements as above. The two lines deterministic and Poisson are indexed by "Dis+TO" to denote the fact that they account for the discrete nature of TCP, and that they use the expressions of Q and Z from [35]. The line for deterministic losses coincides with that of the

[12] One can use another model for the number of packets lost upon congestion if statistics on the loss process at the packet level are available.

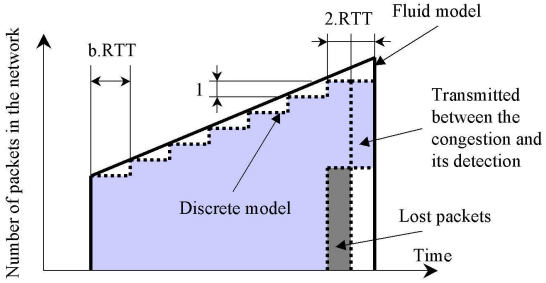

Fig. 11. Fluid model vs. packet model

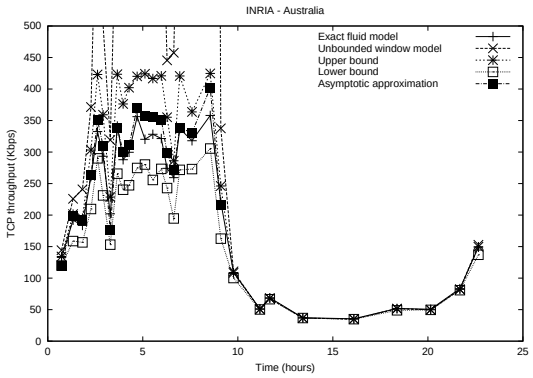

Fig. 12. Long-distance connection

packet level model for all three connections. Both lines are close to the real throughput on the SD and MD connections, but not on the LD one. On the MD connection, the result for deterministic losses is good since the linear rate increase assumption is correct and losses are quite deterministic. On the SD connection, the result is good due to the TCP sub-linearity phenomenon we have explained above. Concerning the result for Poisson losses, it is better than the result for deterministic losses on the LD connection since the loss process on this connection is closer to Poisson than to deterministic. On the MD connection, the real throughput is somewhere between the deterministic and Poisson lines. This is expected since the loss process itself on the MD connection is less variable than a Poisson process.

D. Validation of bounds

To validate the derived bounds for the case of window size limitation, we choose to work on the LD connection where our model for TCP is most appropriate (the rate is linear). We plot the results for the whole day in Fig. 12. First, we see that from 0 to 10 o'clock (slack periods) the throughput calculated in the case of no limit on the congestion window (Eq. (6)) significantly deviates from the exact fluid model. Our exact fluid model accounts for the rate limitation. During the rest of the day both throughputs coincide. This deviation means that the receiver advertised window is frequently reached during the slack periods, and hence our bounds can be applied to estimate the throughput. The saturation of the throughput is clearly reflected by the last line in Table III, where we measure the fraction of time during which the LD connection is transmitting at its maximum window size.

We plot our two bounds given by Proposition 8 on the same figure. They are quite close to the exact fluid model throughput. The figure also shows the throughput obtained by the asymptotic approximation in (32), which we recall is valid when the maximum window size is frequently reached. The figure shows that (32) approximates well the exact fluid model.

IV. Conclusions

In this work, we presented an analysis of TCP throughput under a general loss process. The only assumption we made on the loss process is stationarity and ergodicity. We provided an explicit expression for the throughput in the case of no limit on the transmission rate. The throughput was shown to be inversely proportional to RTT and to the square-root of the packet loss probability, as was already proved in the literature for much simpler loss models [24], [28], [35]. We also provided bounds on the throughput for the case when a limit exists on the maximum window size. Furthermore, we extended our work to include the Timeout mechanism and to account for the discrete nature of TCP. We explained how our model can be used to compute moments of the TCP transmission rate of order higher than 1, in particular the variance.

The importance of our model is justified by the different types of loss processes we observed while measuring Internet traffic. The model we proposed is able to capture any correlation and any distribution of inter-loss times. Several existing models can be seen as particular cases of our general approach. On paths where TCP transmission rate increases linearly between congestion events, our model gives excellent results. However, on paths where TCP window growth is sub-linear, we notice some overestimation of the real throughput. In a future work, we will try to account for this sub-linearity in TCP modeling. We will also try to characterize the loss process at the packet level and to compute based on that, good expressions for the functions Q and Z that model Timeouts. Another future work will be the use of the expression of the variance of the TCP transmission rate to design congestion control mechanisms for real time applications that are friendly with TCP and that exhibit low rate variability.

Acknowledgments

We would like to thank our colleagues at ESSI, ENST and the University of South Australia for providing the required material to conduct our experimentations. We thank Thomas Bonald for his input on the generalization of the square root formula. We also acknowledge the interesting and valuable suggestions made by the anonymous SIGCOMM'2000 and IEEE/ACM ToN referees.

References

[1] M. Allman, H. Balakrishnan, S. Floyd, "Enhancing TCP's Loss Recovery Using Limited Transmit", *RFC 3042*, Jan. 2001.
[2] E. Altman, K. Avratchenkov and C. Barakat, "TCP in presence of bursty losses", *ACM SIGMETRICS*, June 2000.
[3] E. Altman, K. Avranchenkov, C. Barakat, "A stochastic model for TCP/IP with stationary random losses", *ACM SIGCOMM*, August 2000.
[4] E. Altman, K. Avrachenkov, C. Barakat, R. N. Queija, "State-dependent M/G/1 Type Queueing Analysis for Congestion Control in Data Networks", *IEEE INFOCOM*, April 2001.

[5] E. Altman, K. Avrachenkov, C. Barakat, R. N. Queija, "TCP modeling in the presence of nonlinear window growth", *in proceedings of ITC-17*, Septembre 2001.

[6] S. Asmussen and G. Koole, "Marked point processes as limits of Markovian arrival streams", *J. Appl. Prob.*, Vol 30, pp 365-372, 1993.

[7] F. Baccelli and P. Bremaud, "Elements of queueing theory: Palm-Martingale calculus and stochastic recurrences", *Springer-Verlag*, 1994.

[8] C. Barakat, "TCP modeling and validation", *IEEE Network*, vol. 15, no. 3, pp. 38-47, May 2001.

[9] C. Barakat, A. Al Fawal, " Analysis of link-level hybrid FEC/ARQ-SR for wireless links and long-lived TCP traffic", *Performance Evaluation Journal*, vol. 57, no. 4, pp. 423-500, August 2004.

[10] R. Bellman, "Introduction to matrix analysis", McGraw-Hill, New York, 1960.

[11] A. A. Borovkov and S. G. Foss, "Stochastically recursive sequences and their generalizations", *Siberian Advances in Mathematics*, **2**, No. 1, pp. 16-81, 1992.

[12] A. Brandt, "The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients", *Adv. Appl. Prob.*, Vol 18, pp 211-220, 1986.

[13] K. Fall, S. Floyd, "Simulation-based Comparisons of Tahoe, Reno, and SACK TCP", *ACM Computer Communication Review*, vol. 26, no. 3, pp. 5-21, Jul. 1996.

[14] W. Fischer and K. Meier-Hellstern, "The Markov-modulated Poisson process (MMPP) cookbook", *Performance Evaluation*, Vol 18, pp 149-171, 1992.

[15] S. Floyd, "Connections with Multiple Congested Gateways in Packet-Switched Networks Part 1: One-way Traffic", *ACM Computer Communication Review*, Octobre 1991.

[16] S. Floyd and K. Fall, "Promoting the Use of End-To-End Congestion Control in the Internet", *IEEE/ACM Transactions in Networking*, August 1999.

[17] S. Floyd, M. Handley and J. Padhye, "Equation-based congestion control for unicast applications: the extended version", *ACM SIGCOMM*, August 2000.

[18] P. Glasserman and D.D. Yao, "Stochastic vector difference equations with stationary coefficients", *J. Appl. Prob.*, Vol 32, pp 851-866, 1995.

[19] J.J. Hunter, "On the moments of Markov renewal processes", *Adv. Appl. Prob.*, Vol 1, pp 188-210, 1969.

[20] ITU-R, "Allowable Error Performance for a Hypothetical Reference Digital Path Operating at or above the Primary Rate", Recoommendation S.1062-1, 1994-1995.

[21] V. Jacobson, "Congestion avoidance and control", *ACM SIGCOMM*, August 1988.

[22] S.H. Kang and D.K. Sung, "A Markovian arrival process (MAP) modeling for superposed ATM traffic", manuscript.

[23] A. Kumar, "Comparative performance analysis of versions of TCP in a local network with a lossy link", *IEEE/ACM Transactions on Networking*, August 1998.

[24] T.V. Lakshman and U. Madhow, "The performance of TCP/IP for networks with high bandwidth-delay products and random loss", *IEEE/ACM Transactions on Networking*, June 1997.

[25] R. Loynes, "The stability of a queue with non-independent inter-arrival and service times", *Proc. Cambr. Phil. Soc.* **58**, No. 3, pp. 497-520, 1962.

[26] D.M. Lucantoni, K.S. Meier-Hellstern, and M.F. Neuts, "A single-server queue with server vacations and a class of non-renewal arrival processes", *Adv. Appl. Prob.*, Vol 22, pp 676-705, 1990.

[27] M. Mathis, J. Mahdavi, S. Floyd, A. Romanow, "TCP Selective Acknowledgment Options", *RFC 2018*, Oct. 1996.

[28] M. Mathis, J. Semke, J. Mahdavi, and T. Ott, "The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm", *ACM Computer Communication Review*, July 1997.

[29] A. Misra, T. Ott, and J. Baras, "The Window Distribution of Multiple TCPs with Random Queues", *IEEE GLOBECOM*, Decembre 1999.

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

[31] J. Nagle, "Congestion control in IP/TCP internetworks", *RFC 896*, January 1984.

[32] M.F. Neuts, "A versatile Markovian point process", *J. Appl. Prob.*, Vol 16, pp 764-779, 1979.

[33] M.F. Neuts, "Structured stochastic matrices of M/G/1 type and their applications", Marcel Dekker, New York, 1989.

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

[35] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, "Modeling TCP throughput: A simple model and its empirical validation", *ACM SIGCOMM*, Septembre 1998.

[36] LBNL's `tcpdump` tool, available at http://www-nrg.ee.lbl.gov/

[37] S. Savari and E. Telatar, "The Behavior of Certain Stochastic Processes Arising in Window Protocols", *IEEE GLOBECOM*, Decembre 1999.

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

[39] K. Thompson, G.J. Miller, R. Wilder, "Wide-Area Internet Traffic Patterns and Characteristics", *IEEE Network*, vol. 11, no. 6, pp. 10-23, Novembre 1997.

[40] M. Vojnovic, J-Y Le Boudec, "On the Long-Run Behavior of Equation-Based Rate Control", *ACM SIGCOMM*, August 2002.

[41] Y. Zhang, N. Duffield, V. Paxson, S. Shenker, "On the Constancy of Internet Path Properties", *ACM SIGCOMM Internet Measurement Workshop*, San Francisco, California, USA, November 2001.

PLACE PHOTO HERE

Eitan Altman received the B.Sc. degree in electrical engineering (1984), the B.A. degree in physics (1984) and the Ph.D. degree in electrical engineering (1990), all from the Technion-Israel Institute, Haifa. In (1990) he further received his B.Mus. degree in music composition in Tel-Aviv university. Since 1990, he has been with INRIA (National research institute in informatics and control) in Sophia-Antipolis, France. His current research interests include performance evaluation and control of telecommunication networks and in particular congestion control, wireless communications and networking games. He is in the editorial board of several scientific journals: Stochastic Models, JEDC, COMNET, SIAM SICON and WINET. He has been the (co)chairman of the program committee of several international conferences and workshops (on game theory, networking games and mobile networks).

PLACE PHOTO HERE

Konstantin Avrachenkov received the master degree in Control Theory (1996) from St. Petersburg State Technical University and Ph.D. degree in Mathematics (2000) from University of South Australia. Currently he is a research fellow at INRIA Sophia Antipolis, France. His main research interests are Markovchains, Markov decision processes, singular perturbation theory, mathematical programming and the performance evaluation of data networks.

PLACE PHOTO HERE

Chadi Barakat is a permanent research scientist in the Planete research group at INRIA - Sophia Antipolis since March 2002. He got his Electrical and Electronics engineering degree from the Lebanese University of Beirut in 1997, and his master and Ph.D. degrees in Networking from the University of Nice - Sophia Antipolis in 1998 and 2001. His Ph.D. has been done in the Mistral group at INRIA - Sophia Antipolis. From April 2001 to March 2002, he was with the LCA department at EPFL-Lausanne for a post-doctoral position, and from March to August 2004 he was a visiting faculty member at Intel Research Cambridge. Chadi Barakat was the general chair of PAM 2004 and serves in the program committees of many international conferences as Infocom, PAM, WONS, ASWN and Globecom. His main research interests are congestion and error control in computer networks, the TCP protocol, voice over IP, wireless LANs, Internet measurement and traffic analysis, and performance evaluation of communication protocols.