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 flow control mechanism. The rate at which data is transmitted increases linearly in time until a packet loss is detected. At that point, the transmission rate is divided by a constant factor. Losses are generated by some exogenous random process which is only assumed to be stationary. 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 congestion 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 which make simple assumptions on the inter-loss time process.

1. INTRODUCTION

We analyze in this paper the performance of TCP (Transmission Control Protocol), the widely-used transport protocol of the Internet [15, 30]. 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 time 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 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 buffer, etc. (see [5] for a survey on TCP issues).

[*]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 (French National Research Network in Telecommunications) "Constellations" project on satellite communications.

A remarkable attention has been given to TCP modeling within the research community [1, 10, 17, 18, 21, 22, 23, 27, 29]. This is not surprising since 95% of today Internet traffic is carried over TCP. Closed form expressions for the throughput have been obtained. These expressions have helped to understand the impact of different network and TCP parameters on the throughput of the connection and on the efficiency of network resource utilization (e.g. Fairness). 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 [11, 12].

The mathematical analysis of TCP requires two steps. First, we need to construct a model for the window size evolution. Most of the existing models ignore the Slow Start phase and make the assumption that the source stays always in Congestion Avoidance mode. A fluid model has often been used. The window of the connection is assumed to increase linearly as a function of time until a loss occurs; then it is divided by two. An initialization to one packet has been proposed in [23] in case of losses detected via TimeOut. The phase of recovery from losses is assumed to be negligible and the source resumes its linear increase directly after the reduction. In [27], a packet-level model has been proposed to account for the discrete nature of TCP. Indeed, the volume of data in the network at any moment is in multiple of packets. It increases by one once the increase in the window 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 time between congestion events (i.e. between moments at which the congestion window is reduced). Namely, one needs to model the impact of the path between the source and the destination on the connection. Particular models are considered in the literature. The fixed point approach used in [18, 21] gives a deterministic time between congestion events. The assumptions made in [27] can also be shown to imply a deterministic time between congestion events. An exponentially distributed time with a constant intensity has been considered in [23] between congestion events. In [29], the intensity of the exponentially distributed time between congestion events is assumed to increase with the window size. Instead of working in real time, the authors in [22, 26] chose to work in a virtual time. This time is obtained by sampling the congestion window at the moments of ACK arrivals. Again, 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 are found in this virtual time and a method is suggested to go back to the real time. An expression of TCP throughput is provided for each of these models. Our experimentations over the Internet show however that the times between congestion events can have general distribution. They can vary from an approximately deterministic case to a considerably bursty case. Moreover, some correlation may exist between these times. We note, in particular, that if packets are dropped independently with a constant probability then the times between drops are not independent (since the instantaneous transmission rate is variable). We believe that the Internet is so heterogenous that different types of distributions of times between congestion events will always exist. We also believe that making simple assumptions on the process of congestion events will lead to a wrong estimation of TCP throughput. In particular, we have shown in [1] that the throughput of TCP increases when the moments at which the congestion window is reduced tend to appear in bursts. This results in performance problems in applications based on explicit expressions for TCP throughput. Consider for example the case of TCP-friendly applications using explicit expressions for TCP throughput to adapt their rate (see [11, 12] and references therein). These are typically real time applications designed to compete fairly for the available bandwidth with TCP transfers. Suppose that these applications use an expression for TCP throughput that makes a simple assumption on the process of congestion events (e.g. deterministic or Poisson). These applications will suffer when the process of congestion events is highly bursty. They can perform better and transmit at a higher rate by basing their throughput calculation on a more precise model for congestion events.

In this paper, we investigate the case of a general sequence of times between congestion events. In the sequel, we call a congestion event a *loss event*. A loss (event) is a moment where the congestion window is reduced by a given constant factor. It 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) [27]. All we assume is that this process of loss moments is stationary ergodic. With this minimal assumption, we are able to obtain explicit expressions for the throughput of TCP. Our loss model is general enough to allow us to capture any correlation or any distribution of inter-loss times.

For the dynamics of TCP, we decided to focus on the transmission rate which is the number of packets in the network (or the volume of data) divided by the RTT of the connection. The source is assumed to have always data to send. The number of packets in the network is thus equal to the number of packets that can be fit within the window. Denote by $X(t)$ the transmission rate of the connection at time t. At any moment, we can multiply $X(t)$ by RTT to get the window size in terms of packets. We assume that $X(t)$ increases linearly with time at a rate α. If we denote by b the number of data packets covered by one ACK, then $\alpha = 1/(bRTT^2)$. Let ν denote the decrease in the transmission rate when a loss event occurs. Usually ν is equal to one half but we consider a more general scenario to account for other possible flow control mechanisms (see e.g. [6]). The moments of losses are modeled by a general stationary ergodic point process [4] with non-null and finite intensity λ. Let $\{T_n\}_{n=-\infty}^{+\infty}$ be a particular realization of the point process. Consider for the moment the case where losses are quickly detected without the need for a long TimeOut period (e.g. via Duplicate ACK mechanism or via an efficient fine-granularity TimeOut mechanism). Then, the evolution of the transmission rate can be described by the following equation

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

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 [4].

In the next section, we use the machinery of stochastic processes to study this model of the rate evolution. We compute the throughput, that is the time average of process $X(t)$. We also compute the first two moments of the transmission rate at loss arrivals for the stationary regime. Then different examples of loss processes are studied: deterministic, Poisson, i.i.d. and Markovian arrival processes. The expression of the throughput is provided for each of these particular cases. In Section 2.3, we extend our model to account for the case where there is a limitation on the evolution of the transmission rate caused by the receiver advertised window; we provide bounds on the throughput for that model. In Section 2.4, 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 3, we present the testbed as well as the results of our experimentations. In particular, our 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 an exponential distribution. Our experimentations show also the common problem of linear rate increase models. When the transmission rate of TCP indeed exhibits a linear increase, our model gives excellent results. The linear rate increase is known to hold for TCP connections in which the propagation delay is large in comparison with queuing delays, since in that case, RTT is almost constant (see [2]). However, in the case where queueing delays are non negligible as in Local Area Networks, TCP window growth exhibits some sub-linearity due to the increase of the RTT with the window size. In this case, we find that linear rate models overestimate the real throughput. We conclude Section 3 with a method to correct the error caused by the fluid assumption. With this method, the deterministic case of our fluid model has given the same results as the packet level approach described in [27]. Finally, we conclude the work in Section 4.

2. THE MAIN RESULTS

We first note that the equation (1) is a particular case of stochastic linear difference equations [8, 13]. Since the inter-loss time process is stationary ergodic and since $\nu < 1$ and $E[S_n] < \infty$, it follows from Theorem 2A in [13] (see the appendix of [1] for more details) that equation (1) has a stationary solution given by

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

Moreover, if the window evolution starts from an arbitrary window size X_0, it will converge almost sure to the above stationary regime [8, 13]

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

2.1 The computation of the first two moments of X_n and the throughput

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

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

$$E[X_n^*] = \frac{\alpha}{\lambda(1-\nu)} \tag{3}$$

$$E[(X_n^*)^2] = \frac{\alpha^2}{1-\nu^2}[R(0) + 2\sum_{k=1}^{\infty} \nu^k R(k)] \tag{4}$$

Proof. To calculate (3) and (4), we use the expression (2) for the stationary regime.

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

Similarly, we obtain

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

Then, we regroup the terms of the last series to get

$$\begin{aligned}
E[(X_n^*)^2] &= \alpha^2 R(0) \sum_{j=0}^{\infty} \nu^{2j} + 2\alpha^2 \sum_{k=1}^{\infty} R(k)\nu^k \sum_{j=0}^{\infty} \nu^{2j} \\
&= \frac{\alpha^2}{1-\nu^2}[R(0) + 2\sum_{k=1}^{\infty} \nu^k R(k)]. \qquad \square
\end{aligned}$$

Remark 1. *The expectation computed in (3) is taken with respect to the loss instants. This expectation is also referred to as Palm expectation in the context of point processes [4].*

Remark 2. *We note the remarkable insensitivity property, that $E[X_n^*]$ does not depend on the correlation between inter-loss times nor on their moments of order greater than one.*

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

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

Our main result is the following close expression for $\overline{X}$ as a function of the loss intensity λ, the correlation functions $R(k)$, the linear increase factor α (thus b and RTT) and the multiplicative decrease factor ν.

Proposition 2. The throughput is given by

$$\overline{X} = \lambda\alpha[\frac{1}{2}R(0) + \sum_{k=1}^{\infty} \nu^k R(k)] \tag{5}$$

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

$$E[X(t)] = \lambda E^0[\int_0^{T_1} X(t)dt] \tag{6}$$

where $E^0[\cdot]$ is an expectation associated with Palm distribution. In particularly, $P^0\{T_0 = 0\} = 1$. Now using formula (6) and expression (2), we can write

$$\begin{aligned}
E[X(t)] &= \lambda E^0[\int_0^{T_1} (\nu X_0 + \alpha t)dt] = \lambda E^0[\nu X_0 S_0 + \frac{\alpha}{2} S_0^2] \\
&= \lambda E^0[\alpha\nu \sum_{k=0}^{\infty} \nu^k S_{-1-k} S_0] + \frac{\lambda\alpha}{2} E^0[S_0^2] \\
&= \lambda\alpha \sum_{k=0}^{\infty} \nu^{k+1} R(k+1) + \frac{\lambda\alpha}{2} R(0) \\
&= \lambda\alpha[\frac{1}{2}R(0) + \sum_{k=1}^{\infty} \nu^k R(k)] \qquad \square
\end{aligned}$$

Remark 3. *Often the covariance function $C(k) = R(k) - E[S_n]^2$ is used instead of the correlation function $R(k)$. Then, the formulas (4) and (5) become*

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

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

We express now the throughput as a function of the probability that a TCP packet is lost, or more precisely as a function of the probability that a TCP packet causes the source to enter the Fast Recovery phase and to divide its window. Denote this probability by p. This factor is often used in the literature to model the impact of the network on TCP dynamics. Define $A(t)$ as the number of packets transmitted on the TCP connection until time t, and $L(t)$ as the number of loss events until time t. p is simply equal to

$$p = \lim_{t\to\infty} \frac{L(t)}{A(t)} = \lim_{t\to\infty} \frac{\lambda t}{\int_0^t X(\tau)d\tau} = \frac{\lambda}{\overline{X}} = \frac{1}{d\overline{X}}. \tag{7}$$

This allows us to write our main result (5) in another form so as to grasp the influence of p and of RTT on the throughput for general distribution of inter-loss times. Define the normalized correlation function: $\hat{R}(k) = R(k)/d^2$. Then using (7) 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 $\hat{C}(k) = C(k)/d^2$ (where $C(k)$ is defined in Remark 2) then we obtain the following formula for TCP throughput

$$\overline{X} = \frac{1}{RTT\sqrt{pb}}\sqrt{\frac{1+\nu}{2(1-\nu)} + \frac{1}{2}\hat{C}(0) + \sum_{k=1}^{\infty}\nu^k\hat{C}(k)}. \quad (8)$$

As it was already shown for simple loss models [18, 21, 27], we see that also for general losses the throughput of TCP is inversely proportional to RTT and to the square root of p. The distribution and the correlation of inter-loss times figure in the terms below the root.

REMARK 3. *Note that formula (5) can also be rewritten in terms of the second moment of the transmission rate at loss instants*

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

2.2 Examples of loss process

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

2.2.1 IID random losses (General Renewal Process)

Here, 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. Then the formulas (3), (4) and (5) take the following form.

PROPOSITION 3. Let $\{S_n\}_{n=-\infty}^{+\infty}$ be i.i.d. with $d := E[S_n]$ and $d^{(2)} := E[S_n^2]$. Then,

$$E[X_n^*] = \frac{\alpha d}{1-\nu},$$

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

$$\overline{X} = E[X(t)] = \frac{\alpha}{d}[\frac{1}{2}d^{(2)} + \frac{\nu d^2}{1-\nu}]. \quad (9)$$

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

$$\overline{X} = \frac{\alpha d}{1-\nu}. \quad (10)$$

For $\nu = 0.5$, this is similar to the expression for the throughput obtained in [23]. If the inter-loss times are deterministic, we get

$$\overline{X} = \frac{1+\nu}{2(1-\nu)}\alpha d \quad (11)$$

Substituting d from equation (7) into equation (11) and recalling that α is equal to $1/(bRTT^2)$, we get for $\nu = 0.5$ the famous square root formula obtained in the literature [18, 21, 27]

$$\overline{X} = \frac{1}{RTT}\sqrt{\frac{3}{2bp}}, \quad (12)$$

where p is the probability that a TCP packet is lost. This can also be obtained directly from (8) as $\hat{C}(k) = 0$ for all k in the case of deterministic inter-loss times.

REMARK 4. *We note from (9) that the throughput can be expressed as a constant, that depends only on the average inter-loss times, plus a term which grows linearly with the variance of the inter-loss times. Hence for large burstiness in losses (which implies large variance of inter-loss times) we shall get a large underestimate of the throughput if we use the exponentially distributed assumption on inter-loss times (e.g. [23]) or if we assume that they are fixed constants.*

2.2.2 Correlated losses modeled as a Markovian Arrival Process

In this section we consider correlated losses which are modeled by Markovian Arrival Process (MAP) [19, 20, 24, 25]. It was shown in [3] 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, PH-renewal process [25] and Markov Modulated Poisson Process (MMPP) [9] are particular case 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 $\{(i,j)|i \ge 0, 1 \le j \le m\}$ 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 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 $D_{ij}/(-C_{ii} - D_{ii})$.

Let $\{S_n\}_{n=1}^{\infty}$ be a sequence of the inter-arrival times for MAP and let $\{J_n\}_{n=1}^{\infty}$ be a sequence of the states of the underlying Markov process at the arrival epochs. Then $\{J_n, S_n\}_{n=1}^{\infty}$ is a Markov renewal process [14] with the following transition probability matrix [25]

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

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. Now if we take the initial distribution of the underlying Markov chain $J(t)$ as μ, the arrival process becomes interval-stationary or event-stationary. The event-stationary version of MAP has the following joint distribution function for the inter-arrival times [16]

$$F_{S_0 \cdots S_n}(x_0, ..., x_n) = \mu \prod_{i=0}^{n} \{(I - \exp\{Cx_i\})T\}e \qquad (13)$$

Consequently, the joint Laplace-Stieltjes transform is given by

$$f(z_0, ..., z_n) = E[\exp\{-\sum_{k=0}^{n} z_k S_k\}] = \mu \prod_{k=0}^{n} \{(z_k I - C)^{-1}D\}e. \qquad (14)$$

Next, using the Laplace-Stieltjes transform (14), we can easily calculate the first two moments and the correlation function of the inter-arrival time process. Namely,

$$E[S_n] = -\frac{d}{ds}(\mu(zI - C)^{-1}De)|_{z=0} = -\mu C^{-1}e, \qquad (15)$$

$$E[S_n^2] = \frac{d^2}{ds^2}(\mu(zI - C)^{-1}De)|_{z=0} = 2\mu C^{-2}e \qquad (16)$$

$$\begin{aligned} R(k) &= E[S_n S_{n+k}] = \frac{\partial^2}{\partial z_0 \partial z_k} f(z_0, ..., z_k)|_{z_i=0} \\ &= \mu C^{-2} D T^{k-1} C^{-2} De. \end{aligned} \qquad (17)$$

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: $(A^{-1}(z))' = -A^{-1}(z)A'(z)$ $A^{-1}(z)$ [7].

Note that MAP becomes MMPP with infinitesimal generator R and arrival rate matrix Λ, if we take $C = R - \Lambda$ and $D = \Lambda$.

Now, employing (15), (16) and (17), we can calculate the first two moments of the process $\{X_n\}$.

Proposition 4. Let the loss process $\{S_n\}$ be represented by MAP. Then,

$$E[X_n^*] = -\frac{\alpha}{1-\nu}\mu C^{-1}e \qquad (18)$$

$$E[(X_n^*)^2] = \frac{\alpha^2}{1-\nu^2} 2\mu (C^{-2} + \nu C^{-2} D[I - \nu T]^{-1} C^{-2} D)e \qquad (19)$$

Proof. The above formulas are immediately obtained from (3), (4) with the help of (15), (16), (17) and the following derivation

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

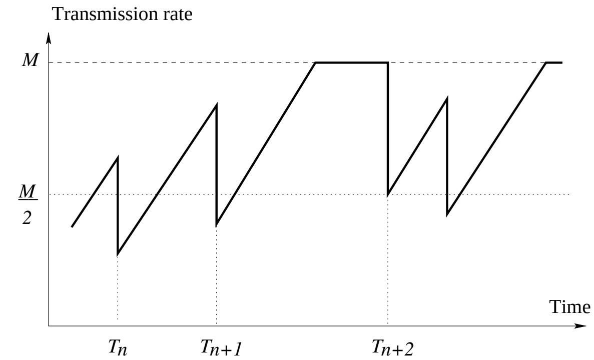

Figure 1: TCP rate evolution with limitation

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

□

Next, using (5), we calculate the throughput of TCP.

Proposition 5. Let $\{S_n\}$ be a Markovian Arrival Process. Then, the throughput of TCP is given by

$$\overline{X} = -\frac{\alpha}{\mu C^{-1} e}\mu(C^{-2} + \frac{1}{2}C^{-2}D[I - \nu T]^{-1}C^{-2}D)e. \qquad (20)$$

2.3 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 indicates the maximum number of packets that can wait at the destination before being handed to the application [23, 30]. Thus, in this section, we shall assume that the transmission rate is limited by a maximal value of M. We take $\nu = 0.5$ given that 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.

The example of the window evolution is presented in Figure 1. The stochastic difference equation (1) is respectively modified to the following form

$$X_{n+1} = M \wedge (\frac{1}{2}X_n + \alpha S_n), \qquad (21)$$

where $\wedge$ stands for *minimum* operation. Note that the model becomes nonlinear and it is probably not possible to obtain the explicit expressions for $E[X_n]$ and $\overline{X}$ for the general loss process. We shall thus use the results of the previous sections to obtain bounds for the performance measures in this case.

Before deriving the bounds we shall establish a stability result for the process X_n using a Loynes-type construction.

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 (21). Furthermore, for any initial state X_0 we have P-a.s.

$$\lim_{n\to\infty} \sup_{k\geq n} |X_k - X_k^*| = 0. \tag{22}$$

That is, for any initial state X_0, the process $\{X_k\}_{k\geq n}$ converges in distribution to the stationary process as $n \to \infty$.

Proof. Define on the same probability space the family of processes $\{X_k^{(n)}, k \in Z\}$, $n = 0, 1, ...$ as follows. $X_k^{(n)} := 0$ for $k \leq -n$, and for $k > -n$ it is given by the recursion (21). For each k, $X_k^{(n)}$ is increasing with respect to n and thus it has a limit (which is obviously finite) which we denote by X_k^*. This limit satisfies (21) since for every n, $X_k^{(n)}$ satisfies it. Finally, the stationarity of the sequence $\{S_n\}$ implies that $\{X_k^*\}$ is stationary as well. The convergence of X_n to X_n^* follows from the fact that

$$\begin{aligned} |X_{n+1} - X_{n+1}^*| &= |M \wedge (\frac{1}{2}X_n + \alpha S_n) - M \wedge (\frac{1}{2}X_n^* + \alpha S_n)| \\ &\leq \frac{1}{2}|X_n - X_n^*| \end{aligned}$$

To show that the above inequality holds, one needs to consider four cases. If the both values of $(\frac{1}{2}X_n + \alpha S_n)$ and $(\frac{1}{2}X_n^* + \alpha S_n)$ are less or alternatively greater than M, then the inequality is obvious. Let us consider then not so obvious cases. For example, let $(\frac{1}{2}X_n + \alpha S_n) > M$ and $(\frac{1}{2}X_n^* + \alpha S_n) < M$, then

$$\begin{aligned} |X_{n+1} - X_{n+1}^*| &= M - (\frac{1}{2}X_n^* + \alpha S_n) \\ &\leq (\frac{1}{2}X_n + \alpha S_n) - (\frac{1}{2}X_n^* + \alpha S_n) \\ &\leq \frac{1}{2}|X_n - X_n^*|. \end{aligned}$$

Thus,

$$|X_n - X_n^*| \leq 2^{-n}|X_0 - X_0^*| \to 0 \quad \text{as} \quad n \to \infty.$$

Since both X_0 and X_0^* are finite (they are bounded between 0 and M) this implies (22).

□

Now we note that equation (21) can be rewritten as

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

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

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

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

$$\check{X}_{n+1} = \frac{1}{2}\check{X}_n + \frac{M}{2} \wedge (\alpha S_n) = \frac{1}{2}\check{X}_n + \alpha(\frac{M}{2\alpha} \wedge S_n), \tag{23}$$

and

$$\hat{X}_{n+1} = \frac{1}{2}\hat{X}_n + M \wedge (\alpha S_n) = \frac{1}{2}\hat{X}_n + \alpha(\frac{M}{\alpha} \wedge S_n). \tag{24}$$

Proposition 6. Let $\{S_n\}$ be a stationary stochastic point process. Assume that $X_0 = \check{X}_0 = \hat{X}_0$. Then for all $n \geq 0$, $\check{X}_n \leq X_n \leq \hat{X}_n$. Moreover,

$$2\alpha\check{d} \leq E[X_n] \leq 2\alpha\hat{d} \tag{25}$$

where the expectation $E[X_n]$ is taken with respect to the stationary regime, $\check{d} = E[\check{S}_n]$, $\hat{d} = E[\hat{S}_n]$, 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 $\check{X}_n \leq X_n$. It holds for $n = 0$. Assume it holds for $n = k$. Then, consider two cases $S_k \leq \frac{M}{2\alpha}$ and $S_k > \frac{M}{2\alpha}$. For $S_k \leq \frac{M}{2\alpha}$, one has

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

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

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

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

Since by the results of [13] 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.

The upper bound is obtained in the similar manner by using the auxiliary process (24).

□

Next we calculate the lower and upper bounds for the throughput.

Proposition 7. Let $\check{d}^{(2)} := E[(\check{S}_n)^2]$, $\check{R}(k) := E[\check{S}_{n-k}S_n]$ and $\hat{d}^{(2)} := E[(\hat{S}_n)^2]$, $\hat{R}(k) := E[\hat{S}_{n-k}S_n]$. Then, the lower and upper bounds for the throughput are given by

$$\overline{X} \geq \alpha\lambda\left(\check{R}(0) - \frac{1}{2}\check{d}^{(2)} + \sum_{k=0}^{\infty}\frac{1}{2^{k+1}}\check{R}(k+1)\right), \tag{26}$$

$$\overline{X} \leq \alpha\lambda\left(\hat{R}(0) - \frac{1}{2}\hat{d}^{(2)} + \sum_{k=0}^{\infty}\frac{1}{2^{k+1}}\hat{R}(k+1)\right). \tag{27}$$

Proof. To obtain the lower bound for the throughput we again use the auxiliary process (23). Suppose that $\{X_n\}$ is in a stationary regime and define

$$\check{X}(t) = \begin{cases} \frac{1}{2}\check{X}_n^* + \alpha t, & t \in [T_n, \check{T}_n], \\ \frac{1}{2}\check{X}_n^* + \alpha\check{S}_n, & t \in [\check{T}_n, T_{n+1}], \end{cases} \tag{28}$$

where $\check{T}_n = T_n + \check{S}_n$ (See Figure 2). Similarly to (2), one can write the expression for the stationary version of $\{\check{X}_n\}$, that is

$$\check{X}_n^* = \alpha\sum_{k=0}^{\infty}(\frac{1}{2})^k\check{S}_{n-1-k}.$$

Now, using (28) and the above expression for $\check{X}_n^*$, we obtain

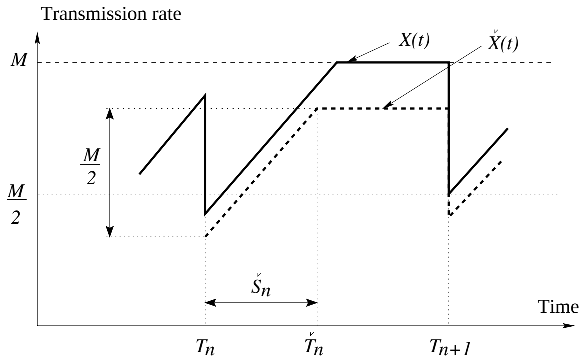

Figure 2: A lower bound for the transmission rate

the lower bound

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

Now, by using the auxiliary process (24), one can calculate the upper bound for the throughput in the similar fashion. □

Note that the two bounds given in Proposition 7 coincide with the throughput given by (5) as $M/(\alpha d) \to \infty$. However, when $M/(\alpha d) \to 0$, the upper bound provided in (27) goes to $2M$. Therefore, we propose to take as an upper bound the minimum between M and the upper bound given in (27). The lower bound converges to the following expression

$$\overline{X} \simeq M - \frac{\lambda M^2}{8\alpha} \tag{29}$$

as $M/(\alpha d)$ goes to zero. As was shown in [1, 27], the expression (29) appears to be a very good approximation of the throughput when the maximal rate is frequently reached. It is the throughput obtained by TCP when the transmission rate reaches its maximum between each two losses.

2.3.1 Bounds for Poisson and IID cases

Assume first that the loss process is Poisson. Then formulas (26) and (27) 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 throughput for a Poisson loss process in the case of an infinite maximum window size (see Subsection 2.2.1).

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,

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

2.3.2 Bounds for MAP loss process

Note that the quantities $\check{d}, \hat{d}$, $\check{d}^{(2)}, \hat{d}^{(2)}$ and $\check{R}(k), \hat{R}(k)$ can also be easily calculated for the loss process modeled by MAP. Indeed,

$$\begin{aligned}
\check{d} &= E[\check{S}_n] = \int_0^{\infty} (\frac{M}{2\alpha} \wedge x)\mu dF(x)e \\
&= \mu \int_0^{\infty} (\frac{M}{2\alpha} \wedge x) \exp\{xC\} dx\, De \\
&= \mu[I - \exp\{\frac{M}{2\alpha}C\}]C^{-2}De,
\end{aligned}$$

$$\hat{d} = E[\hat{S}_n] = \mu[I - \exp\{\frac{M}{\alpha}C\}]C^{-2}De,$$

and, similarly,

$$\check{d}^{(2)} = \mu\left(-\frac{M}{\alpha}\exp\{\frac{M}{2\alpha}C\}C^{-2}D + 2[\exp\{\frac{M}{2\alpha}C\} - I]C^{-3}D\right)e,$$

$$\hat{d}^{(2)} = \mu\left(-2\frac{M}{\alpha}\exp\{\frac{M}{\alpha}C\}C^{-2}D + 2[\exp\{\frac{M}{\alpha}C\} - I]C^{-3}D\right)e.$$

To compute $\check{R}(k)$ and $\check{R}(k)$, $k \geq 1$, we use the joint distribution (13).

$$\begin{aligned}
\check{R}(k) &= \int_0^{\infty} \cdots \int_0^{\infty} (\frac{M}{2\alpha} \wedge x_0)x_k\ \mu \exp\{Cx_0\}D \cdots \\
&\qquad \exp\{Cx_k\}De\ dx_0 \cdots dx_k \\
&= \mu[I - \exp\{\frac{M}{2\alpha}C\}]C^{-2}DT^{k-1}C^{-2}De \\
\hat{R}(k) &= \mu[I - \exp\{\frac{M}{\alpha}C\}]C^{-2}DT^{k-1}C^{-2}De
\end{aligned}$$

For $k = 0$, we have

$$\check{R}(0) = \int_0^{\infty} (\frac{M}{2\alpha} \wedge x)x\mu \exp\{Cx\}Dedx =$$

$$\mu\left(-\frac{M}{2\alpha}\exp\{\frac{M}{2\alpha}C\}C^{-2}D + 2[\exp\{\frac{M}{2\alpha}C\} - I]C^{-3}D\right)e$$

$$\hat{R}(0) =$$

$$\mu\left(-\frac{M}{\alpha}\exp\{\frac{M}{\alpha}C\}C^{-2}D + 2[\exp\{\frac{M}{\alpha}C\} - I]C^{-3}D\right)e$$

2.4 Modeling conservative TimeOuts

We assumed till now that losses are quickly detected. This is indeed the case when losses are detected via duplicate ACKs (the Fast Retransmit algorithm) or via a fine-granularity correctly-set retransmission timer. However, most TCP implementations use a coarse-granularity timer for the detection of losses in the case where three duplicate ACKs cannot be received. The objective is to be conservative and to reduce the load on the operating system. 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 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).

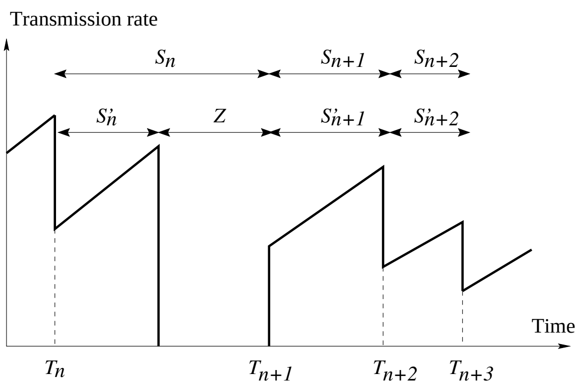

Figure 3: A model for TCP with TO and TD losses

The idle period involves the time between 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. We include in this section these idles times into our explicit expression for the throughput. Similarly, it can be included into the expressions for the bounds on the throughput.

Let λ_{TO} denote the number of TO losses per unit of time. Let Z denote the average duration of the idle period after TO losses. 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 $T_n - Z$ until instant T_n where the transmission is resumed. During the idle period, the transmission rate is equal to zero. 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 and can be neglected. A problem with this assumption is that it is not valid when the retransmission timer is backed-off. After a back-off of the retransmission timer, the source gets in the linear increase phase at a small window. However, our experimentations have shown that this back-off is rare. We depict in Figure 3 a sample of the transmission rate evolution in presence of TO losses according to our model.

Define now the sequence (see Figure 3)

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

PROPOSITION 8. Assume that the process S'_n is stationary [1] and let $\lambda' = 1/E[S'_n]$ and $R'(k) = E[S'_n S'_{n+k}]$. Then the throughput is given by

$$\overline{X} = \lambda' \alpha [\frac{R'(0)}{2} + \sum_{k=1}^{\infty} \frac{1}{2^k} R'(k)](1 - \lambda_{TO} Z). \qquad (30)$$

PROOF. Consider a new point process $\{T'_n\}$ whose inter-arrival times are $\{S'_n\}$, and consider the process describing the evolution of the transmission rate when we eliminate the idle periods. This is shown in Figure 4. This is precisely the model described by equation (1), where S_n is replaced by S'_n and T_n replaced by T'_n. According to Proposition 2, the throughput for the new model is given by

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

This is equal to the number of packets transmitted in the *original model* averaged only over the periods other than TimeOuts.

The expected number of TO losses that occur during any interval $[0, t]$ is $\lambda_{TO} t$. Hence the fraction of time in which the connection is idle in the original model is given by $\lambda_{TO} Z$. The throughput in the original model is thus $(1 - \lambda_{TO} Z)\overline{X}' + \lambda_{TO} Z \cdot 0$ which establishes (30).

□

[1] 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 $\{S'_n\}$ of being stationary.

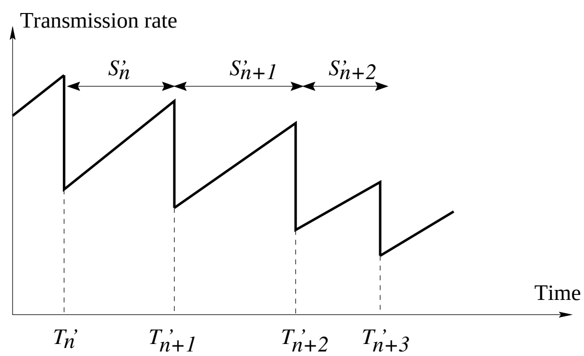

Figure 4: A model for TCP with TO and TD losses when eliminating the TimeOut periods

λ_{TO} and Z are not a part of the network model but rather the results of a particular reaction of the TCP source to packet losses. We provide here a method for the approximation of these two parameters as a function of times between losses. The calculation is inspired by [27]. Assume that we are able to characterize the process $\{S'_n\}$ and thus to calculate $\overline{X}'$. The problem is how to infer λ_{TO} and Z in order to calculate the real throughput $\overline{X}$. Define Q as the probability that a loss is of type TO. Q is related to our parameters by the equation $\lambda_{TO} = Q\lambda$. In [27], Q and Z are calculated as functions of the probability that a TCP packet is lost. More precisely, it is the probability that a packet causes the source to enter the Fast Recovery phase and to divide its window by two. Denote this probability by p. As in Section 2.2.1, p can be taken equal to $\lambda'/\overline{X}'$. Using the expressions of Q and Z from [27], the calculation of $\overline{X}$ is straightforward.

3. MODEL VALIDATION

Our work has been mainly motivated by the fact that the loss of TCP packets over the Internet may present a more complicated structure than the simplistic approaches proposed in the existing literature. We have run real TCP connections between several Internet sites. For each connection we measure the time between loss events (i.e. between moments at which the window is divided by two). 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 introduce at the end 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 [27]. This can also be considered as a proof that [27] accounts only for deterministic loss processes.

3.1 Experimentation testbed

We ran at different days during the month of January 2000, three 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 south of France. The destination machines are located respectively at the ESSI school at 1 km from INRIA (`nessie.essi.fr`), at the ENST school at Paris (`solo.enst.fr`), and at the University of South Australia (`linus.levels.unisa.edu.au`). TCP Packets are of 1460 Byte size (not including 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. These can be considered as three long-life TCP transfers across LAN (Local Area Network), MAN (Metropolitan Area Network), and WAN (Wide Area Network) networks. The three destinations implement the Delay ACK mechanism, so $\alpha = 1/(2RTT^2)$.

Data packets and the corresponding ACKs are captured with the `tcpdump` tool [28] at INRIA. Given the TCP version of the source, we developed a tool that looks at the trace of every connection and provides the instants of window reductions (T_n). In the case of a loss detected via TimeOut, the tool provides the duration of the TimeOut period. In general, our tool determines the times S_n and S'_n, the parameters λ_{TO} and Z, and the average RTT from the measured trace file.

Each connection is run for 24 hours. Its trace however is divided at fixed intervals in a way to get a sufficient number of losses. We fix this interval to 20 minutes for ESSI, to 40 minutes for ENST and to one hour for Australia. This gives us a set of trace files for every connection. For each trace file, we measured the real throughput of the connection and we compared it to the expected throughput using both the expressions we derived from our fluid model, as well as the expression from the packet level model in [27]. We examined the validity of different models for the distribution of inter-loss times (deterministic, Poisson, iid, general correlated). The different moments and correlation functions of inter-loss times are calculated from the trace file.

We studied separately the effect of assumptions on the loss process and the correctness of our fluid model for TCP 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 a linear increase and multiplicative decrease with a silence time during TimeOuts and with the maximum limit on the congestion window; see Figure 8). We call this process the *Exact Fluid Model* and we calculate exactly its throughput. This throughput is calculated by computing the area below the transmission rate between two losses, and then by summing all the areas and dividing the result by the transfer time. The exact fluid model

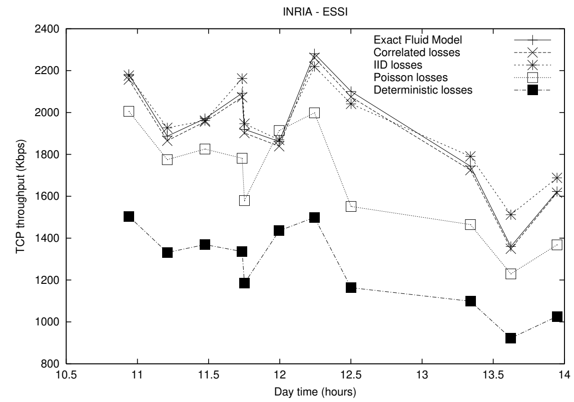

Figure 5: LAN connection

throughput is exactly the throughput we are trying to estimate in our analysis and which we (and other authors that use linear rate increase models to study TCP) are claiming represents real TCP throughput. Our measure of how well a given model for the distribution of inter-loss times approximates the real loss process will be how close the throughput predicted by our close form expression (5) agrees with the exact fluid model throughput. If the loss model is good (according to the above criterion), we are still not guaranteed that the measured throughput of the real TCP connection agrees with our throughput formulas. We expect the latter to be close to real TCP throughput if the linear rate increase model is appropriate, which is not always the case as we will see later. Ideally, the exact fluid model throughput must not deviate from the real TCP throughput and the throughput calculated under a given assumption on the type of the loss process should follow closely that of the exact fluid model.

3.2 Validation of the model for losses

We focus in this section on the identification of the loss process. As already mentioned, our measure of how well a given model for the distribution of inter-loss times approximates the real loss process, is how close the throughput predicted by our close form expression (5) agrees with the exact fluid model throughput. Different loss processes are considered: deterministic, Poisson, general iid, general correlated.

We plot in Figures 5,6, and 7 the results for the three connections (LAN,MAN, and WAN respectively) as a function of time. 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 (5). We notice also that the iid model gives approximately the same result as the correlated model which means that losses are rarely correlated especially in MAN and WAN networks. Some correlation can be seen in the LAN environment. For instance, around 13:30 the loss process has a negative correlation coefficient equal to -0.2.

Consider now the Poisson and deterministic cases. We look in Figures 5,6, and 7 at points where losses are iid. It is clear from the expression of the throughput in the iid case (9) that at a constant loss intensity, the throughput of TCP increases with the variance of inter-loss times. Thus, a com-

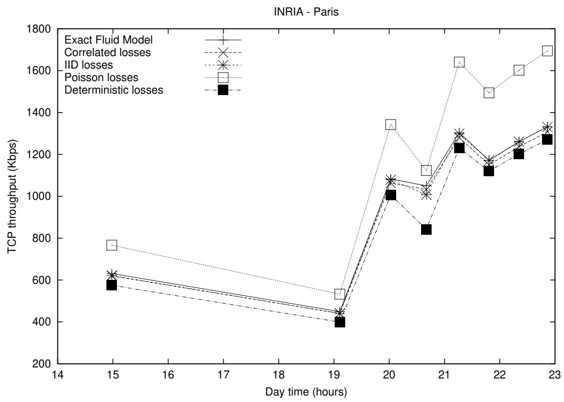

Figure 6: MAN connection

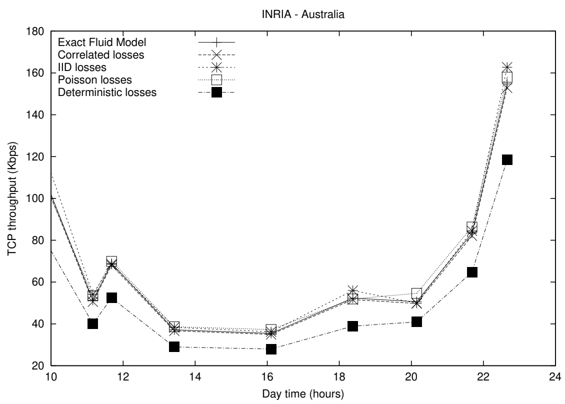

Figure 7: WAN connection

parison of the exact fluid model throughput to the throughput predicted in the Poisson and deterministic cases indicate how much the times between losses vary. For example, the further we are from the throughput predicted in the deterministic case, the more important is the variance of times between losses. On the LAN 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 Figure 8. On such connection, one expects that models assuming deterministic inter-loss times will give bad results compared to the real throughput. On the MAN 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 WAN connection, it is clear that losses occur according to a Poisson process.

The above three figures show well that the inter-loss time process change from one path to another. It changes also on the same path. This observation confirms the utility of our general model for losses on TCP connections.

3.3 Validation of the model for TCP

In this section we compare the exact fluid model to real TCP on the three connections we are considering. Our objective here is to test the validity of the TCP modeling (the linear growth assumption as well as the fluid assumption). The results are provided in Figures 9, 10, and 11. On the LAN and

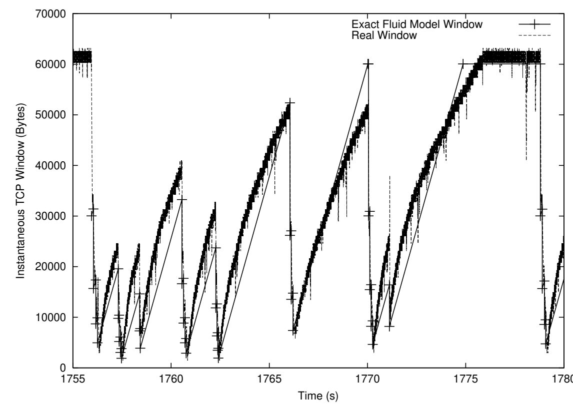

Figure 8: Fluid model vs. real window on the LAN connection

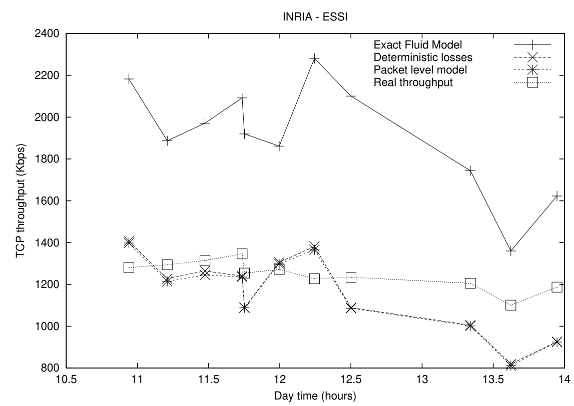

Figure 9: LAN connection

MAN connections, the exact fluid model overestimates real TCP. This overestimation increases with the real throughput. It is mainly due to the sub-linearity of TCP congestion window evolution, which can be seen in Figure 8. This sub-linearity is in turn due to the increase of the RTT with the congestion window (i.e. due to the increase of queuing time with respect to propagation delay). Linear rate increase models assume that the RTT is independent of the window size and use the average of the RTT to calculate the slope of the transmission rate increase. This overestimation does not exist on the WAN connection where the window is usually small and where the propagation delay plays a major role.

Models taking simple assumptions about the loss process deviate from the exact fluid model which in turn deviates from real TCP. Occasionally, they can give good results when the two deviations cancel each other. As an example, the assumption that inter-loss times are deterministic works well on the LAN connection, even though the loss process has a considerable variation (see Figure 8). This is due to the cancellation effect we described. Our approach solves the problem of inter-loss time modeling. However, the problem of TCP window evolution modeling still persists.

Another source for overestimation in our model is the fluid approximation. In fact, the transmission rate of TCP does not increase continuously but rather jumps when the number

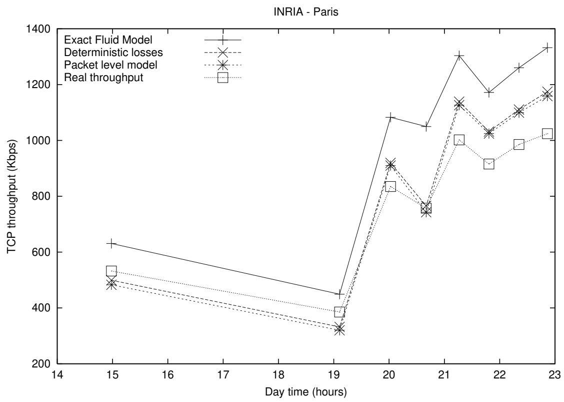

Figure 10: MAN connection

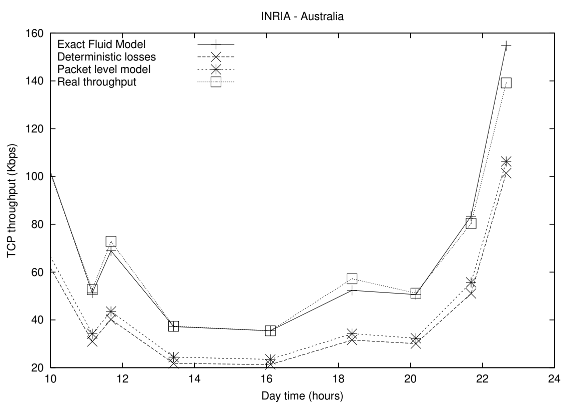

Figure 11: WAN connection

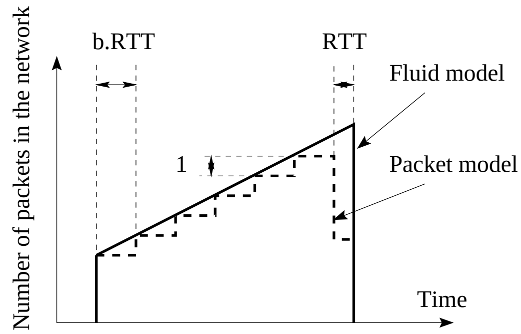

Figure 12: Fluid model vs. packet model

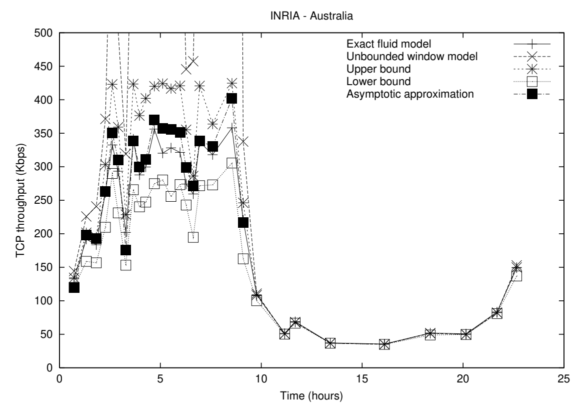

Figure 13: WAN connection

of packets injected into the network increases by one. However, the window at the source can be assumed to change continuously as a function of time. We use the packet level model in [27] to correct this problem of discreteness. Figure 12 shows how the number of packets in the network changes as a function of time between two congestion events. The dashed thick line represents our fluid model and the continuous thick line represents the discrete TCP behavior described in [27]. 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 deterioration caused by the last RTT before the resumption of the transmission (Figure 12). It has been assumed in [27] that the number of packets transmitted during the last RTT is equal to half the window size. Thus, half $E[X_n^*] * RTT$ must be subtracted from the average integral of the transmission rate between two loss events in order to account for this last RTT before the resumption of the transmission. We get the following throughput

$$\begin{aligned}\overline{X}_d &= \overline{X} - \frac{1}{2RTT} - \frac{E[X_n^*] * RTT}{2E[S_n]} \\ &= \overline{X} - \left(\frac{b}{2} + \frac{1}{2(1-\nu)}\right)\alpha RTT\end{aligned}$$

Using this correction, we compare in Figures 9, 10, and 11 our model with deterministic inter-loss times (11) with the packet level model in [27]. The results of these two models practically coincide for all three connections. These results are close to the real throughput on the LAN and MAN connection and not on the WAN one. In general, the model in [27] is close to the real throughput in two cases: when the linear rate increase assumption is correct and losses are quite deterministic (MAN connection) or when the linear rate increase assumption is not valid and inter-loss times vary considerably (LAN connection). The correctness of the model in the latter case is due to error cancellation between the model for TCP and the model for losses.

3.4 Validation of bounds

To validate the derived bounds for the case of window size limitation, we choose to work on the WAN connection where our model for TCP is most appropriate. We plot the results for the whole day (Figure 13). First, we see that from 0 to 10 o'clock the throughput calculated in the case of no limit on the congestion window significantly deviates from the exact fluid model. Our exact fluid model accounts for the rate limitation. This deviation means that the receiver advertised window is frequently reached during these hours and our bounds can be applied to estimate the throughput. We plot our two bounds on the same figure. They are indeed quite close to the exact fluid model throughput. We plot also the throughput obtained by the asymptotic approximation in equation (29). This approximation is valid only when the maximum window is frequently reached. The figure shows that it approximates well the exact fluid model. One can envisage the use of this approximation once the expectation of the transmission rate at loss moments computed by (3) exceeds the maximum rate.

4. CONCLUSIONS

In this work, we presented an analysis of TCP throughput under a general loss process. The only assumption on the loss process under consideration is stationarity and ergodicity. We provide explicit expression for the throughput in the case of no limit on the transmission rate. The throughput is shown to be inversely proportional to RTT and to the square root of the packet loss probability, as was already observed for much simpler loss models [18, 21, 27]. We further provided bounds for the case when a limit exists on the maximum window size. We extend our work to include also the TimeOut mechanism.

The importance of our model is justified by the different types of loss processes we observed when measuring Internet traffic. The model we proposed is able to capture any correlation or 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 indeed increases linearly between congestion moments, 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. Another direction for future research could be the identification of the parameters of the Markovian Arrival Process proposed to characterize losses on correlated paths.

5. ACKNOWLEDGMENTS

We would like to thank our colleagues at ESSI, ENST and the University of South Australia who provided us with the required material to conduct our experimentations. We thank Thomas Bonald for his suggestion to add the relation between our model with deterministic losses and the square root formula. We acknowledge also the interesting and valuable suggestions made by the anonymous SIGCOMM referees.

6. REFERENCES

[1] E. Altman, K. Avrachenkov, and C. Barakat, "TCP in presence of bursty losses", *ACM SIGMETRICS*, Jun 2000.

[2] 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", *34th IEEE Conference on Decision and Control*, Dec 1995.

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

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

[5] C. Barakat, E. Altman, and W. Dabbous, "On TCP Performance in a Heterogeneous Network : A Survey", *IEEE Communications Magazine*, Vol 38, No 1, pp 40-46, Jan 2000.

[6] C. Barakat, N. Chaher, W. Dabbous, and E. Altman, "Improving TCP/IP over Geostationary Satellite Links", *IEEE GLOBECOM*, Dec 1999.

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

[8] 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.

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

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

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

[12] S. Floyd, M. Handley, J. Padhye, and J. Widmer, "Equation-based congestion control for unicast applications", *ACM SIGCOMM*, Aug 2000.

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

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

[15] V. Jacobson, "Congestion avoidance and control", *ACM SIGCOMM*, Aug 1988.

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

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

[18] 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*, Jun 1997.

[19] 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.

[20] D.M. Lucantoni, "New results on the single server queue with a batch Markovian arrival process", *Commun. Statist. - Stoch. Models*, Vol 7, pp 1-46, 1991.

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

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

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

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

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

[26] 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.

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

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

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

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

[31] M. Zorzi and R. Rao, "Effect of correlated errors on TCP", *Proceedings of CISS'97*, 1997.