A Model of TCP in Wide Area Networks

Sophie Fortin and Bruno Sericola
IRISA-INRIA
Campus de Beaulieu
35042 Rennes cedex, France
{sophie.fortin}{bruno.sericola}@irisa.fr

Abstract

This paper presents a discrete-time Markov chain model for the Reno version of TCP, the transmission control protocol for reliable transport on the Internet. The purpose is the evaluation of stationary TCP flows behavior using performance measures such as the mean throughput. The model is based on previous works which are generalized by taking into account the slow start phases that appear after each time-out recovery, which importance is discussed.

1. Introduction

Because of the great expand of the Internet, a lot of work has been done on its efficiency and on possible improvements. The apparently simple mechanism of the *Transport Control Protocol* TCP used by HTTP transfer, file transfer, email and remote access has been modeled with various stochastic tools.

Assuming a periodic window evolution marked by random loss events of probability p, separating successive congestion avoidance phases, the authors of [15] have shown that the mean throughput ρ was $O(1/\sqrt{p})$.

Many studies are based on a fluid approach and are usually and mainly interested in getting an analytical expression for the mean throughput of a single steady-state TCP connection. It is the case of [14], [13], [16] and [6], but also [2], [1], [5] and [4] which focus on the window size W_n just before the n-th loss. The case of multiple TCP connections is the subject of [3], [12], and [8] for instance. Among all other tools explored, the max-plus algebra provides in [7] expressions for the mean throughput in the case of several routers in series.

Our paper is based on previous works presented in [19], [18] and [9] which consider a discrete-time model and a discrete evolution of the window size. We propose here a discrete-time Markov chain model which aims to give analytical expressions for measures such as the mean throughput of one bulk transfer TCP-Reno flow among exogenous traffic. A flow may represent the transfer of a large data file as well as the global TCP traffic from one ftp serveur to another for instance. This model also provides various results for the successive TCP phases.

The paper is organized as follows. The TCP-Reno mechanisms are reviewed in Section 2 and modeled in Section 3 with a discrete-time Markov chain based on the notion of *rounds*. An expression for the mean throughput is obtained in Section 4 in which we also discuss our numerical results. Section 5 shows the importance of slow start phases in terms of duration and of number of segments sent. Section 6 concludes the paper.

2. Description of TCP

TCP is a reliable flow control protocol for connection oriented links (see [21] and [10]). Network congestion, identified by packets loss, is detected by TCP by the arrival or not of packet acknowledgments, leading the protocol to a modification of the transmission throughput.

Indeed, each successfully transmitted packet is validated and confirmed to the source by a small packet called ACK (*ACKnowledge*) which contains the sequence number of the next expected byte and a receiver's maximum window size giving information about its buffer occupancy. So as not to unnecessarily load the network, the receiver sometimes waits for more data to acknowledge before sending an ACK. Those ACKs are thus called *delayed* ACKs. The number b of segments validated per ACK is typically equal to 1 or 2. A timer T_s will set the departure of an ACK if no new data is to be ACKed.

There are two kinds of loss detection :

- detection by *time-out*, or TO : if no ACK is received for byte number n before the expiry of a timer T_0, then a *time-out* occurs. The segment starting with byte n is considered lost and is thus retransmitted, and no more data is sent until byte n is ACKed;

Proceedings of the 10th IEEE Int'l Symp. on Modeling, Analysis, & Simulation of Computer & Telecommunications Systems (MASCOTS'02)
1526-7539/02 $17.00 © 2002 **IEEE**

- detection by the arrival of *three duplicate* ACKs, or TD : if a segment beginning with byte n is lost but some following segments are received, each of these will generate an ACK requesting byte n, that is one ACK requesting byte n and successive duplicate ACKs. The reception of the third duplicate ACK (4 similar ACKs) will halve the window and generate the segment retransmission. In fact, duplicate ACKs can be due to disordered segment reception, and the arrival of one or two duplicate ACKs is not considered as a proof of loss.

TCP is based on a sliding window dynamic. The window, initialized to 1, gives the number of bytes that can be sent before receiving any ACK. Each time an ACK arrives, the window slides to the right to release into the network as many bytes as the ACK validates. The function of TCP is to modify the window size W^c (in segments) according to the algorithm presented below and described in the RFC2001 ([22]).

First, TCP-Reno consists in three phases depending on loss events and on the comparison of the congestion window size W^c to the *slow start threshold* W^{th}. If a TD loss occurs, then $W^{th} := \max(\lfloor W^c/2 \rfloor, 2)$ and $W^c := \max(\lfloor W^c/2 \rfloor, 1)$, then starts a congestion avoidance phase. If a TO loss occurs, then $W^{th} := \max(\lfloor W^c/2 \rfloor, 2)$, $W^c := 1$, and a time-out phase starts.

Time-out, slow start and congestion avoidance operate as follows

- *time-out* (to) : just after a TO loss detection, the apparently lost segment is retransmitted. After each retransmission failure, the timer value doubles (from T_0 to $2T_0$, $4T_0$, $8T_0$,...) until $64T_0$, and then remains constant (and gets back to T_0 at the end of this time-out period, that is when the corresponding ACK arrives).

- *slow start* (ss) : starts after a time-out recovery and lasts as long as $W^c < W^{th}$. During slow start, $W^c := W^c + 1$ each time an ACK is received (b segments ACKed). If the whole window gets successfully transmitted, then it will generate $\lceil W^c/b \rceil$ ACKs, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. For $b = 1$, a window of size W^c will thus generate W^c ACKs, so it will grow from W^c to $2W^c$. Consequently, the congestion window grows exponentially during the slow start phase;

- *congestion avoidance* (ca) : each ACK reception adds $1/W^c$ segments to the window size, so that the ACKment of the whole window increases W^c by $1/b$. Consequently, the congestion window grows linearly (of one segment every b rounds) during the congestion avoidance phase.

3. The model

If the dispatch duration of all the segments and of all the ACKs held in a given window is negligible compared to the *round trip time* RTT, then we can justify the following definition of *round* given in [19], [18] and [9] : a *round* is the period of time between the departure of the first segment of the current window and the arrival of its ACK. The duration of a round is close to the round trip time when the delayed ACK timer T_s is small compared to the RTT.

3.1. Definition

We aim to model the window behavior using a homogeneous discrete-time Markov chain $X = (X_n)_{n \geq 1}$ with two components $X_n = (W_n^c, W_n^{th})$. The first component W_n^c denotes, when positive, the window size during the n-th round. The null value for W_n^c is used to represent the time-out period. The second component W_n^{th} denotes the value of the slow start threshold during the n-th round. We denote by $W_{\max}$ the maximum window size, which is the receiver's buffer capacity indicated in the ACKs (when W_n^c reaches $W_{\max}$ it remains constant until the next loss). The description the state space of this Markov chain is given, more formally, by

- $X_n = (i, j)$ with $i \in \{1, \ldots, W_{\max}\}$ and $j \in \{2, \ldots, \lfloor W_{\max}/2 \rfloor\}$ when the current window size is i and the slow start threshold is j,

- $X_n = (0, j)$ with $j \in \{2, \ldots, \lfloor W_{\max}/2 \rfloor\}$ when the connection is in a time-out period with threshold j.

As long as $W_n^c = i \geq 1$, a transition of the Markov chain represents one round and thus lasts RTT seconds. In order to make the mean duration (in seconds) of a time-out period $E[T_{to}]$ equal to RTT times the mean number of successive visits to state $(0, j)$, we define the two following transitions from each state $(0, j)$, $j = 2, \ldots, \lfloor W_{\max}/2 \rfloor$:

- from $(0, j)$ to $(1, j)$ with probability p_0 at the end of a time-out period,

- from $(0, j)$ to $(0, j)$ with probability $1 - p_0$ otherwise,

with $p_0 = RTT/E[T_{to}]$. In Subsection 4.5, we give the expression of $E[T_{to}]$ as a function of RTT, p and T_0.

The state space E of this Markov chain is a subset of the set E' defined by $E' = \{0, \ldots, W_{\max}\} \times \{2, \ldots, \lfloor W_{\max}/2 \rfloor\}$. We can notice that for $W_{\max} = 10, 50, 100, 200$, the set E' contains respectively 44, 1224, 4949 and 19899 states.

A simple example of the beginning of a connection is given in Figures 1 and 2 where we take $W_0^{th} = 4$ segments, $W_{\max} = 8$ and $b = 1$.

Proceedings of the 10th IEEE Int'l Symp. on Modeling, Analysis, & Simulation of Computer & Telecommunications Systems (MASCOTS'02)
1526-7539/02 $17.00 © 2002 IEEE

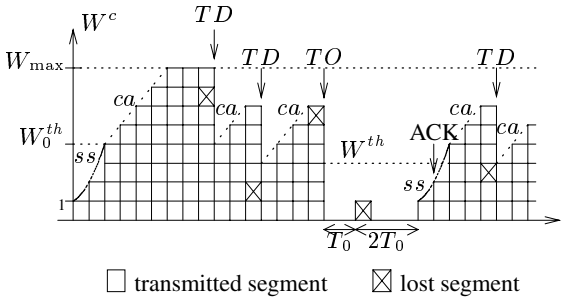

Figure 1. Example of congestion window evolution.

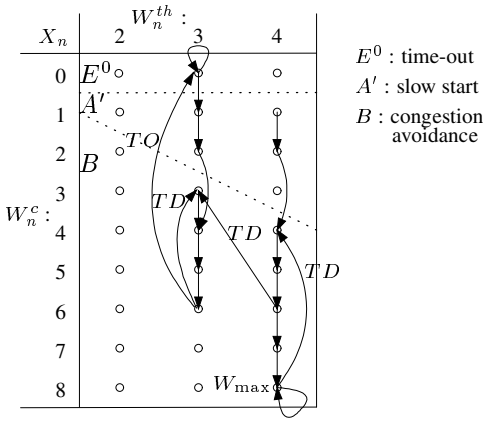

Figure 2. Markov chain transitions and partitioning.

It can be noted in Figure 1, that, for instance, state $(3, 4)$ will never be reached. This is due to the fact that the window sizes reached in the slow start phase are for

- $b = 1 : 1, 1 + \left\lceil \frac{1}{b} \right\rceil = 2, 2 + \left\lceil \frac{2}{b} \right\rceil = 4, 8, 16, 32, \ldots$
- $b = 2 : 1, 1 + \left\lceil \frac{1}{b} \right\rceil = 2, 2 + \left\lceil \frac{2}{b} \right\rceil = 3, 5, 8, 12, \ldots$

This example leads to the following partitioning for the state space of the Markov chain, which is represented in Figure 2. The state space E is written as $E = E^0 \cup A \cup B$ where

- $E^0 = \{(0, j) \mid 2 \leq j \leq \lfloor W_{\max}/2 \rfloor\}$,
- $B = \{(i, j) \mid 2 \leq j \leq i \leq W_{\max}$ and $j \leq \lfloor W_{\max}/2 \rfloor\}$,
- $A = \{(i, j) \mid 1 \leq i < j \leq \lfloor W_{\max}/2 \rfloor$ and $\exists n \geq 0$ such that $i = f^{[n]}(1)\}$, where $f(w) = w + \lceil w/b \rceil$, $f^{[0]}(w) = w$, and $f^{[n]} = f^{[n-1]} \circ f$, for $n \geq 1$.

The partition shown in Figure 2 is in fact a partition of the state space E' and the set A contains the reachable states of A' during the slow start phase.

This discrete-time Markov chain is irreducible and aperiodic. It is thus ergodic and its stationary distribution π is the unique distribution verifying $\pi P = \pi$ where P is the transition probability matrix, which is given in Section 3.3.

3.2. Cyclic behavior

In what follows, we consider the Markov chain in stationary regime and we assume that the source behaves as a saturated source, which means that there are always packets waiting for transmission.

In such a context, an observation of the congestion window size shows a cyclic evolution, consisting in one slow start phase followed by several congestion avoidance phases separated by TD losses, and then a TO loss starting a time-out period at the end of which a new cycle begins (see Figure 3).

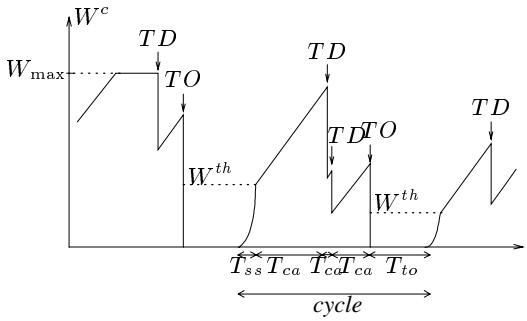

Figure 3. Description of a cycle.

We denote respectively by

- T_{to}, T_{ss} and T_{ca} the duration of a time-out period, a slow start phase and a congestion avoidance phase,
- d_{to}, d_{ss} and d_{ca} the number of segments sent during the periods T_{to}, T_{ss} and T_{ca},
- $T_{E^0}^{back}$ the time between a time-out recovery and the next TO loss,
- $d_{E^0}^{back}$ the number of segments sent during $T_{E^0}^{back}$,
- N_{loss} the mean number of loss detections per cycle,
- ρ the connection throughput, more precisely the mean transmission rate or *send rate*, which takes into account all segments that have left the source, including lost segments and retransmissions (ρ is the input rate seen by the network).

Proceedings of the 10th IEEE Int'l Symp. on Modeling, Analysis, & Simulation of Computer & Telecommunications Systems (MASCOTS'02)
1526-7539/02 $17.00 © 2002 **IEEE**

Observing Figure 3 and because of the cyclic window evolution, we would write the throughput ρ as

$$\frac{E[d_{E^0}^{back}] + E[d_{to}]}{E[T_{E^0}^{back}] + E[T_{to}]}. \quad (1)$$

However, this formula does not take into account the residual rounds that appear after each loss and which are presented in Subsection 4.1 together with the expression of ρ.

3.3. The transition probabilities

We assume that losses only occur in the direction from the sender to the receiver (no loss of ACKs) and that any segment has a fixed probability p to get lost. More precisely, the random variable defined by the number of consecutive segments that are transmitted before loss has a geometric distribution with parameter $1-p$.

Let us first suppose that the connection is in slow start, i.e. $W_n^c = i < j = W_n^{th}$. As long as the Markov chain remains in slow start, the congestion window increases by 1 segment each time an ACK is received. And because $\lceil W_n^c/b \rceil$ segments are acknowledged for the whole round, $W_{n+1}^c = W_n^c + \lceil W_n^c/b \rceil = \lceil \gamma W_n^c \rceil$ with $\gamma = 1 + 1/b$. In the following propositions, we give expressions for the non-zero transition probabilities of the Markov chain. These expressions being easy to obtain, we omit the proofs.

Proposition 1 *For $1 \leq i < j \leq \lfloor W_{\max}/2 \rfloor$, we get :*

- $P_{(i,j)(\lceil \gamma i \rceil, j)} = (1-p)^i$ *: no loss occurs,*
- $P_{(i,j)(0,\max(\lfloor i/2 \rfloor, 2))} = \left(1-(1-p)^i\right) q_i$ *: a TO loss occurs,*
- $P_{(i,j)(\max(\lfloor i/2 \rfloor, 1), \max(\lfloor i/2 \rfloor, 2))} = \left(1-(1-p)^i\right)(1-q_i)$ *: a TD loss occurs,*

where q_i (computed in Subsection 4.2) denotes the probability that a loss is due to time-out when $W^c = i$.

Suppose now that the transmission is in congestion avoidance in state (i,j), i.e. $W_n^c = i \geq j = W_n^{th}$.

Proposition 2 *Observing that congestion avoidance globally raises the window size by $1/b$, i.e. by 1 segment every b rounds, then for $1 \leq j \leq i < W_{\max}$,*

- $P_{(i,j)(i,j)} = (1-p)^i \left(1 - \frac{1}{b}\right)$ *: no loss occurs,*
- $P_{(i,j)(i+1,j)} = (1-p)^i \frac{1}{b}$ *: no loss occurs,*
- $P_{(i,j)(0,\max(\lfloor i/2 \rfloor, 2))} = \left(1-(1-p)^i\right) q_i$ *: a TO loss occurs,*
- $P_{(i,j)(\max(\lfloor i/2 \rfloor, 1), \max(\lfloor i/2 \rfloor, 2))} = \left(1-(1-p)^i\right)(1-q_i)$ *: a TD loss occurs.*

Note that in order to get the model more accurate about the raise of 1 segment every b rounds, we should decompose the Markov chain state (i,j) into b new states, say $(i,j,1)$, $(i,j,2)$, ..., (i,j,b), but, first that would of course significantly increase the Markov chain size (even for $b=2$) and secondly, that would not change the measures of interest since the stationary distribution on the state space E remains the same after such a transformation.

Proposition 3 *Similarly, for each j we have :*

- $P_{(W_{\max},j)(W_{\max},j)} = (1-p)^{W_{\max}}$ *: no loss occurs,*
- $P_{(W_{\max},j)(0,\max(\lfloor W_{\max}/2 \rfloor, 2))} = \left(1-(1-p)^{W_{\max}}\right) q_{W_{\max}}$ *: a TO loss occurs,*
- $P_{(W_{\max},j)(\max(\lfloor W_{\max}/2 \rfloor, 1), \max(\lfloor W_{\max}/2 \rfloor, 2))} = \left(1-(1-p)^{W_{\max}}\right)(1-q_{W_{\max}})$ *: a TD loss occurs.*

As explained in Section 3, we define the transition probabilities in time-out.

Proposition 4 *For each j, we have :*

- $P_{(0,j)(0,j)} = 1 - \frac{RTT}{E[T_{to}]}$ *: no acknowledgment yet,*
- $P_{(0,j)(1,j)} = \frac{RTT}{E[T_{to}]}$ *: the acknowledgment is arrived.*

The expression of $E[T_{to}]$ as a function of the timer T_0 and the loss probability p is computed in Section 4.5.

The shape of the transition probability matrix P and the regions corresponding to the different types of losses are shown in Figure 4.

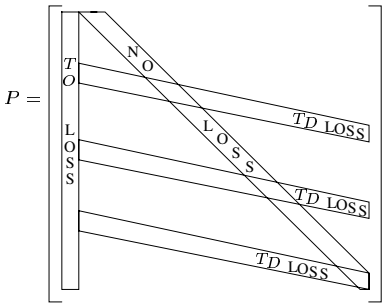

Figure 4. Link between the transition matrix P and TCP.

Proceedings of the 10th IEEE Int'l Symp. on Modeling, Analysis, & Simulation of Computer & Telecommunications Systems (MASCOTS'02)
1526-7539/02 $17.00 © 2002 IEEE

4. Computation of the throughput

4.1. Residual rounds

First, we make the assumption that in a given round, the loss of one segment leads to the loss of the following segments (correlated losses). This should be the case in a high speed network for instance. Moreover, in the round where the loss takes place, if k segments are however transmitted before congestion, then those segments will generate ACKs and the window will slide. This means that k new segments are transmitted in the next round, which is called the $residual\ round$.

This behavior is shown in Figure 5 which depicts the case where the last segment sent during the residual round is lost.

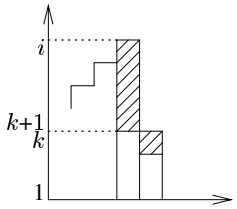

Figure 5. The residual round.

We consequently introduce the following notations :

- d_{rr} : number of segments sent in a residual round,
- p_{rr} : probability that a loss is followed by a residual round, that is probability that a residual round is not empty.

We can now give the expression of the send rate ρ.

Proposition 5 *The send rate ρ is given by*

$$\rho = \frac{E[d_{to}] + E[d_{E^0}^{back}] + N_{loss}E[d_{rr}]}{E[T_{to}] + E[T_{E^0}^{back}] + RTT(N_{loss} - 1)p_{rr}}. \quad (2)$$

Proof. The first terms of expression (2) correspond to Relation (1). The last terms, where N_{loss} appears, are due to the residual rounds. In counting the mean number of segments transmitted during a cycle, we also need to take into account the mean number of segments constituting the residual rounds generated by the N_{loss} loss detections. This mean number of segments is equal to $N_{loss}E[d_{rr}]$.

For what concerns the mean cycle duration, it is increased by $p_{rr}RTT$ for each of the $(N_{loss} - 1)$ TD losses, because the TO loss residual round is taken into account in next time-out period, as shown in Figure 6. □

The expressions of N_{loss}, p_{rr}, $E[d_{rr}]$, $E[d_{to}]$, $E[T_{to}]$, $E[T_{E^0}^{back}]$, $E[d_{E^0}^{back}]$ are given in Relations (4-8) and (11-12).

Remark 1 Let us denote by n_{ca} the number of ca phases in a cycle and $N_{ca} = E[n_{ca}]$. Whereas it is clear that $E[T_{E^0}^{back}] = E[T_{ss}] + E[n_{ca}T_{ca}]$, our numerical results have shown that $E[T_{E^0}^{back}]$ is very closed to $E[T_{ss}] + N_{ca}E[T_{ca}]$, which means that n_{ca} and T_{ca} can be considered as independent. The same results hold for variables n_{ca} and d_{ca}.

4.2. TO-type losses proportion

Now that we introduced residual rounds, we are able to understand how a loss might be a TO loss and not a TD loss, and thus to compute probabilities q_i that a loss is due to TO when $W^c = i$, which are necessary for the evaluation of transition probabilities.

Proposition 6 *The probability q_i that a loss is due to TO when $W^c = i$ is given by : $q_i = 1$ if $i \leq 2b + 1$ and if $i \geq 2b + 2$*

$$q_i = \frac{(1 - (1-p)^{2b+1})\left(1 + (1-p)^{2b+1} - (1-p)^i\right)}{1 - (1-p)^i}. \quad (3)$$

Proof. Using the notation in Figure 5, we have

- If $i \leq 2b + 1$ then $k \leq 2b$ and thus no TD loss can happen (3 duplicate ACKs need $b + b + 1 = 2b + 1$ segments to be received). In this case, the loss is necessarily due to TO, i.e. $q_i = 1$.
- If $i \geq 2b + 2$ then :
  - if $k \leq 2b$: similarly, only a TO loss can occur;
  - if $k \geq 2b + 1$: there is a TO loss only when less than $2b+1$ segments from the residual round arrive at destination (the $2b+1$ first segments are not all received), i.e. the l-th segment from the residual round gets lost, with $1 \leq l \leq 2b + 1$.

  Thus, if we denote by L_{k+1} the event corresponding to the loss of the $(k + 1)$-th segment, we get

$$\begin{aligned} q_i &= P\left(TO \mid W^c = i \ \&\ \text{loss}\right) \\ &= \sum_{k=0}^{i-1} q_{i,k} P\left(L_{k+1} \mid W^c = i \ \&\ \text{loss}\right) \end{aligned}$$

  where

$$\begin{aligned} q_{i,k} &= P\left(TO \mid W^c = i \ \&\ L_{k+1}\right) \\ &= \begin{cases} 1 & \text{if } k \leq 2b \\ 1 - (1-p)^{2b+1} & \text{if } k \geq 2b + 1, \end{cases} \end{aligned}$$

  and $P\left(L_{k+1} \mid W^c = i \ \&\ \text{loss}\right) = \dfrac{(1-p)^k p}{1 - (1-p)^i}$. Relation (3) then follows after some algebra. □

Proceedings of the 10th IEEE Int'l Symp. on Modeling, Analysis, & Simulation of Computer & Telecommunications Systems (MASCOTS'02)
1526-7539/02 $17.00 © 2002 **IEEE**

4.3. Mean number of losses per cycle

Proposition 7 *The mean number N_{loss} of loss detections per cycle is given by*

$$N_{loss} = \frac{1 - \sum_{(i,j)\in E} (1-p)^i \pi(i,j)}{\sum_{(i,j)\in E} q_i \left(1-(1-p)^i\right) \pi(i,j)}. \quad (4)$$

Proof. Each cycle (see Figure 3) is composed of several TD losses and only one TO loss. Thus, we have

$$\frac{1}{N_{loss}} = P(TO \mid \text{loss } \& \ W^c \geq 1) = \sum_{i=1}^{W_{\max}} q_i p_{i|loss}$$

where

$$\begin{aligned} p_{i|loss} &= P(W^c = i \mid \text{loss } \& \ W^c \geq 1) \\ &= \frac{P(\text{loss} \mid W^c = i) P(W^c = i \mid W^c \geq 1)}{P(\text{loss} \mid W^c \geq 1)} \\ &= \frac{\left(1-(1-p)^i\right) \frac{P(W^c=i)}{P(W^c \geq 1)}}{\sum_{i=1}^{W_{\max}} \left(1-(1-p)^i\right) \frac{P(W^c=i)}{P(W^c \geq 1)}} \\ &= \frac{\left(1-(1-p)^i\right) \sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} \pi(i,j)}{\sum_{(i,j)\in E} \left(1-(1-p)^i\right) \pi(i,j)}. \quad \square \end{aligned}$$

4.4. The weight of residual rounds

Proposition 8 *The probability p_{rr} that a residual round appears after loss is given by*

$$p_{rr} = 1 - p \frac{1 - \sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} \pi(0,j)}{1 - \sum_{(i,j)\in E} (1-p)^i \pi(i,j)}. \quad (5)$$

Proof. Let K be the random variable equal to the number of segments sent before loss in the round in which that loss occurred (see Figure 5, in which we have drawn the case $K = k$). We thus have

$$\begin{aligned} p_{rr} &= P(K \neq 0 \mid \text{loss } \& \ W^c \geq 1) \\ &= \sum_{i=1}^{W_{\max}} P(K \neq 0 \mid W^c = i \ \& \ \text{loss}) p_{i|loss} \\ &= \sum_{i=1}^{W_{\max}} \left(1 - \frac{p}{1-(1-p)^i}\right) p_{i|loss}, \end{aligned}$$

which leads to Relation (5) using the expression of $p_{i|loss}$ given in the proof of proposition 7. $\square$

Proposition 9 *The mean number of segments $E[d_{rr}]$ that are sent in a residual round is given by*

$$E[d_{rr}] = \frac{1-p}{p} - \frac{\sum_{(i,j)\in E} i(1-p)^i \pi(i,j)}{1 - \sum_{(i,j)\in E} (1-p)^i \pi(i,j)}. \quad (6)$$

Proof. As above, we denote by K the random variable equal to the number of segments sent before loss in the round in which that loss occurred (see Figure 5). We have

$$E[d_{rr}] = E[K \mid \text{loss } \& \ W^c \geq 1] = \sum_{i=1}^{W_{\max}} E[d_{rr} \mid i] p_{i|loss}$$

where

$$\begin{aligned} E[d_{rr} \mid i] &= E[K \mid W^c = i \ \& \ \text{loss}] = \sum_{k=0}^{i-1} k \frac{(1-p)^k p}{1-(1-p)^i} \\ &= \left(\frac{1-p}{p}\right) \frac{1-(1-p)^i - ip(1-p)^{i-1}}{1-(1-p)^i}. \end{aligned}$$

Relation (6) is then obtained using the expression of $p_{i|loss}$ given in the proof of Proposition 7. $\square$

4.5. Time-out study

The behavior of TCP during a time-out period is illustrated in Figure 6, where rr denotes the residual round (see also Figure 5).

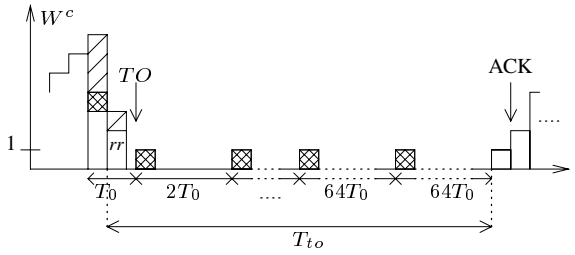

Figure 6. Detail of a time-out period.

The following result can be found in [19].

Proposition 10 *The mean number of segments sent during a time-out period and the mean duration of a time-out period are given by*

$$E[d_{to}] = \frac{p}{1-p} \quad (7)$$

(geometric distribution of segments loss),

$$E[T_{to}] = T_0 \frac{1+p+2p^2+4p^3+8p^4+16p^5+32p^6}{1-p} - RTT. \quad (8)$$

Proceedings of the 10th IEEE Int'l Symp. on Modeling, Analysis, & Simulation of Computer & Telecommunications Systems (MASCOTS'02)
1526-7539/02 $17.00 © 2002 **IEEE**

4.6. Between two time-out periods

In the following remark, we briefly recall some results on sojourn times in Markov chains. These results have been obtained in [20].

Remark 2 Consider an irreducible discrete time Markov chain with finite state space E, transition probability matrix P and stationary probability distribution π. We denote by $\mathbb{1}$ the column vector with all the entries equal to 1. Let F be a proper subset of E and F' the complementary subset $E - F$. The partition F, F' of E induces the following decomposition of P, π and $\mathbb{1}$:

$$P = \begin{pmatrix} P_F & P_{F,F'} \\ P_{F',F} & P_{F'} \end{pmatrix}, \pi = \begin{pmatrix} \pi_F & \pi_{F'} \end{pmatrix}, \mathbb{1} = \begin{pmatrix} \mathbb{1}_F \\ \mathbb{1}_{F'} \end{pmatrix}.$$

If v_i denotes the stationary probability that a sojourn in F initiates in state i ($i \in F$) and v the row vector composed of the v_i, then

$$v = \frac{\pi_F(I - P_F)}{\pi_F(I - P_F)\mathbb{1}_F} = \frac{\pi_{F'}P_{F',F}}{\pi_{F'}P_{F',F}\mathbb{1}_F}, \tag{9}$$

where I is the identity matrix of dimension given by the context. Moreover, for every $i \in F$, let $N_{i,F}$ be the number of visits to state i during a sojourn in F and let r_i be any real number. If we denote by r_F the column vector composed of the r_i and by C_F the random variable $C_F = \sum_{i \in F} r_i N_{i,F}$, we easily get

$$E[C_F] = v(I - P_F)^{-1} r_F = \frac{\pi_F r_F}{\pi_{F'}P_{F',F}\mathbb{1}_F}. \tag{10}$$

Using these results, we have the following proposition.

Proposition 11 *The mean time $E[T_{E^0}^{back}]$ between the end of a time-out period (the beginning of slow start) and the next TO loss is given by*

$$E[T_{E^0}^{back}] = \frac{RTT}{p_0}\left(\frac{1}{\sum_{j=2}^{\lfloor W_{\max}/2 \rfloor}\pi(0,j)} - 1\right). \tag{11}$$

Proof. $E[T_{E^0}^{back}]$ is RTT times the mean time spent by the Markov chain in subset $A \cup B$. Following Remark 2, Relation (10), we have

$$E[T_{E^0}^{back}] = RTT \times E[C_{A\cup B}] = RTT\frac{\pi_{A\cup B} r_{A\cup B}}{\pi_{E^0}P_{E^0,A\cup B}\mathbb{1}_{A\cup B}}$$

where $r_{A\cup B} = \mathbb{1}_{A\cup B}$. We thus have

$$\begin{aligned} E[T_{E^0}^{back}] &= RTT\frac{\sum_{(i,j)\in A\cup B}\pi(i,j)}{\sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} p_0\pi(0,j)} \\ &= \frac{RTT}{p_0}\frac{1 - \sum_{(i,j)\in E^0}\pi(i,j)}{\sum_{j=2}^{\lfloor W_{\max}/2 \rfloor}\pi(0,j)} \\ &= \frac{RTT}{p_0}\frac{1 - \sum_{j=2}^{\lfloor W_{\max}/2 \rfloor}\pi(0,j)}{\sum_{j=2}^{\lfloor W_{\max}/2 \rfloor}\pi(0,j)}, \end{aligned}$$

where the last equality follows from the fact that E^0 is the subset of states $(0,j)$, $j = 2, \ldots, \lfloor W_{\max}/2 \rfloor$. □

Proposition 12 *The mean number $E[d_{E^0}^{back}]$ of segments sent between the end of a time-out period and the next TO loss is given by*

$$E[d_{E^0}^{back}] = \frac{\sum_{(i,j)\in A\cup B} i\pi(i,j)}{p_0\sum_{j=2}^{\lfloor W_{\max}/2 \rfloor}\pi(0,j)}. \tag{12}$$

Proof. $E[d_{E^0}^{back}]$ is the mean number of segments sent during a sojourn in $A \cup B$. We thus have $E[d_{E^0}^{back}] = E[C_{A\cup B}]$ where the entry (i,j) of vector $r_{A\cup B}$ is now the number of segments sent when the Markov chain is in state $(i,j) \in A \cup B$, that is $r_{(i,j)} = i$, for every $(i,j) \in A \cup B$. Following Remark 2, Relation (10), the rest of the proof is similar to that of Proposition 11. □

4.7. Numerical Results

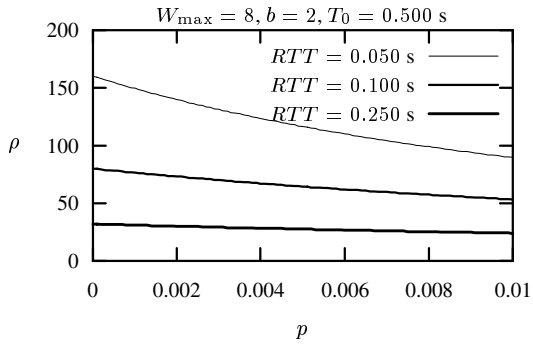

Figure 7. Send rate ρ for different values of RTT.

Figure 7 : The send rate ρ gets equal to $W_{\max}$ segments per RTT ($W_{\max}/RTT$ segments per second) when loss probability p is close to zero, and converges to zero when p increases. Moreover, the shorter RTT is, the more segments are sent per second (quick acknowledgments).

Proceedings of the 10th IEEE Int'l Symp. on Modeling, Analysis, & Simulation of Computer & Telecommunications Systems (MASCOTS'02)
1526-7539/02 $17.00 © 2002 IEEE

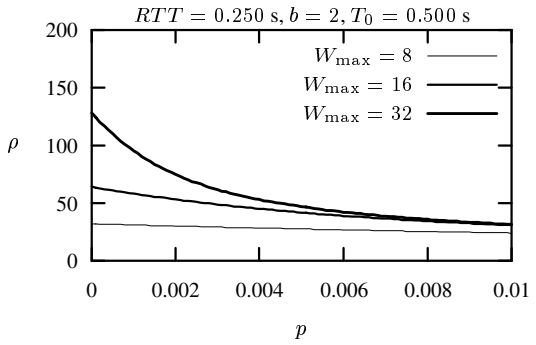

Figure 8. Send rate ρ for different values of $W_{\max}$.

Figure 8 : When $W_{\max}$ increases, the window size can reach higher values and the mean throughput naturally increases too. Note that for small values of the loss probability p, ρ reaches $W_{\max}/RTT$ segments per second, and for large values of p, ρ seems to be less dependent on $W_{\max}$. Indeed, for $p = 0,01$, ρ gets close to 20 or 30 segments per second, that is around 6 segments per RTT for $W_{\max} = 8, 16, 32$.

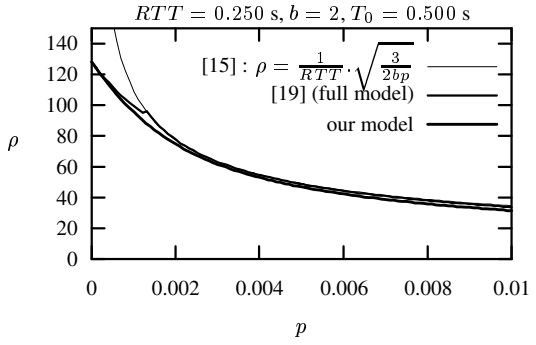

Figure 9. Comparison to other models for $RTT = 0.250$ s.

Figures 9 and 10 provide a comparison to simpler models [15] and [19] which have been validated from both simulations and real traffic measurements that we do not report here but that can be found in [15, 19, 9]. Note that the throughput of our model, evaluated with less simplifications, is lower than the one obtained by the authors of [15] and [19]. But the higher RTT is, the closer the different models are.

Remark 3 Note that similar results have been obtained in [11] for the *goodput* (or *output rate*), which represents the throughput seen by the receiver.

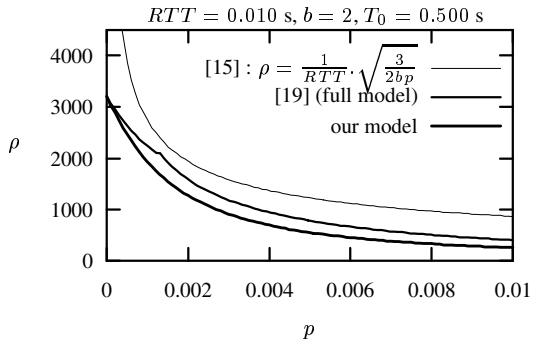

Figure 10. Comparison to other models for $RTT = 0.010$ s.

5. The importance of slow start

The strength of our model is that it allows us to give a detailed description of the window evolution. In particular, we obtain the expression of $E[T_{ss}]$, the mean duration of a slow start phase, and of $E[d_{ss}]$, the mean number of segments sent in a slow start phase.

Proposition 13 *The mean duration $E[T_{ss}]$ of a slow start phase is*

$$E[T_{ss}] = \frac{RTT}{p_0} \frac{\sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} \sum_{i=1}^{j-1} \pi(i,j)}{\sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} \pi(0,j)}.$$

Proof. $E[T_{ss}]$ is RTT times the mean time spent by the Markov chain in subset A. Following Remark 2, Relation (10), we have

$$E[T_{ss}] = RTT E[C_A] = RTT \frac{\pi_A r_A}{\pi_{E^0 \cup B} P_{E^0 \cup B, A} \mathbb{1}_A}$$

where $r_A = \mathbb{1}_A$. We thus have

$$\begin{aligned} E[T_{ss}] &= RTT \frac{\sum_{(i,j)\in A} \pi(i,j)}{\sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} p_0 \pi(0,j)} \\ &= RTT \frac{\sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} \sum_{i=1}^{j-1} \pi(i,j)}{p_0 \sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} \pi(0,j)}, \end{aligned}$$ □

Proposition 14 *The mean number $E[d_{ss}]$ of segments sent during a slow start phase is*

$$E[d_{ss}] = \frac{\sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} \sum_{n=1}^{n_j} w_n (1-p)^{d_{n-1}} \pi(0,j)}{\sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} \pi(0,j)},$$

Proceedings of the 10th IEEE Int'l Symp. on Modeling, Analysis, & Simulation of Computer & Telecommunications Systems (MASCOTS'02)
1526-7539/02 $17.00 © 2002 **IEEE**

where in any slow start phase, w_n is the size of n-th round, $d_n = \sum_{k=1}^{n} w_k$ (with $d_0 = 0$) is the number of segments sent during the first n rounds and n_j is the number of rounds needed to reach the slow start threshold j.

Proof. Let us denote by Z_A the state of subset A by which a sojourn in A begins. These states are necessarily the states $(1, j)$ for $j = 1, \ldots, \lfloor W_{\max}/2 \rfloor$. From Remark 2, Relation (9), $P(Z_A = (1, j))$ is equal to the entry $(1, j)$ of the vector $\pi_A(I - P_A)/[\pi_A(I - P_A)\mathbb{1}_A]$, that is

$$\begin{aligned} P(Z_A = (1,j)) &= \frac{[\pi_A(I-P_A)](1,j)}{\pi_A(I-P_A)\mathbb{1}_A} \\ &= \frac{[\pi_{E^0}P_{E^0,A}](1,j)}{\pi_{E^0}P_{E^0,A}\mathbb{1}_A} \\ &= \frac{p_0\pi(0,j)}{\sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} p_0\pi(0,j)}. \end{aligned}$$

Now, if the slow start phase initiates by state $(1, j)$ then the maximum number of rounds in that phase is equal to n_j. For $n < n_j$, the w_n segments of the n-th round are sent if no loss has occurred during the $n - 1$ first rounds, that is among the d_{n-1} first segments. Thus

$$E[d_{ss} \mid Z_A = (1,j)] = \sum_{n=1}^{n_j} w_n (1-p)^{d_{n-1}}.$$

The result follows by writing

$$E[d_{ss}] = \sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} E[d_{ss} \mid Z_A = (1,j)] P(Z_A = (1,j)). \quad \square$$

We can notice in Figure 11 that the proportion of time spent in slow start per cycle depends on $W_{\max}$ in the sense that when $W_{\max}$ gets higher, slow start phases can reach higher thresholds and thus last longer (whereas in congestion avoidance, the bigger the window size is, the higher is the probability that a loss appears, stopping thus the congestion avoidance phase). But the main remark is that the duration of a slow start phase may reach 10 or 15 per cent of $E[T_{E^0}^{back}]$.

Contrary to slow start duration, Figure 12 shows that the number of segments $E[d_{ss}]$ sent in slow start remains less than 5 per cent of $E[d_{E^0}^{back}]$, even for a high $W_{\max}$. This implies that in the expression of ρ given in Relation 2, the numerator will not change a lot if slow start is not taken into account, but the denominator will be significantly reduced, and thus ρ may significantly grow.

The best way of neglecting slow start phase is to consider that this phase is instantaneous. So if we denote by ρ' the throughput obtained without integrating slow start phases, we have

$$\rho' = \frac{E[d_{to}] + (E[d_{E^0}^{back}] - E[d_{ss}]) + N_{loss}E[d_{rr}]}{E[T_{to}] + (E[T_{E^0}^{back}] - E[T_{ss}]) + RTT(N_{loss} - 1)p_{rr}}.$$

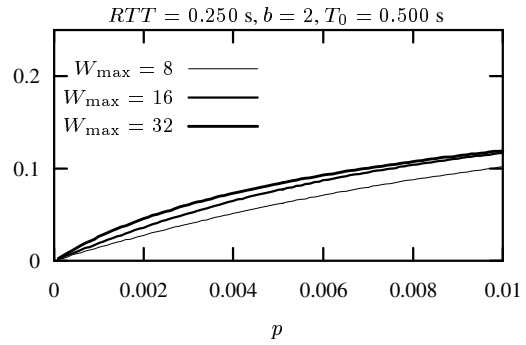

Figure 11. Proportion of time in each cycle : $E[T_{ss}]/E[T_{E^0}^{back}]$.

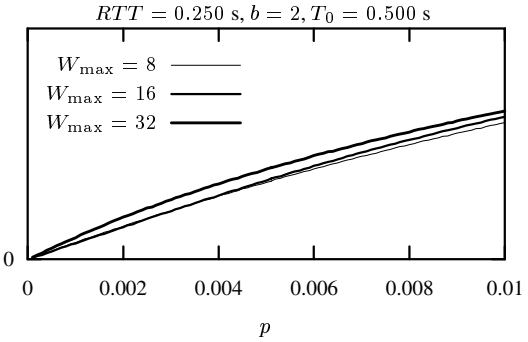

Figure 12. Number of segments in each cycle : $E[d_{ss}]/E[d_{E^0}^{back}]$.

Figure 13, shows that ρ' can be until 12 per cent higher than ρ. The lower the loss probability p is, the more ρ' is close to ρ. But traffic management and bandwidth allocation for instance need a good estimation of ρ, and even a 5 per cent overestimation can lead to severe problems in performance evaluation of other measures of interest.

6. Conclusion

The main assumption we made is that the connection is established in a high speed and wide area (large RTT) network. Indeed, the time needed to send all segments in congestion window and the time interval between ACKs must be significantly low compared to the round trip time for the identification of separated bursts, called and defined as *rounds*.

Proceedings of the 10th IEEE Int'l Symp. on Modeling, Analysis, & Simulation of Computer & Telecommunications Systems (MASCOTS'02)
1526-7539/02 $17.00 © 2002 **IEEE**

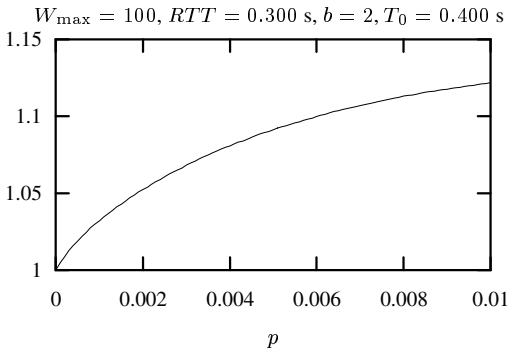

Figure 13. ρ'/ρ versus the loss probability p.

Moreover, we supposed that the loss probability p is independent to the window size, because in high capacity networks, the load of a single connection is not responsible of congestion. Concerning loss correlation (when a segment gets lost, all the following ones in the same round also get lost), we apply our model to high capacity and high speed networks with drop-tail routers, in which the connection is not the cause of congestion and packets of a given round arrive in burst in the overflowed router. And despite multiplexing, a router remains full as long as packets of the same window arrive and thus rejects all of them.

With these assumptions, we have been able to obtain an analytical expression for the throughput of a long term steady-state connection (stationary regime). But our model gives a more precise description of TCP which permits an accurate study of its performance. Other performance measures can be discussed such as, for instance, the proportion of TO-type losses, the average time interval between two consecutive losses, and the proportion of time during which the window size is maximum.

References

[1] A. A. Abouzeid, M. Azizoglu, and S. Roy. Stochastic Modeling of a Single TCP/IP Session over a Random Loss Channel. In *DIMACS workshop on Mobile Networks and Computing*, Rutgers University, New Jersey, US, March 1999.

[2] A. A. Abouzeid, S. Roy, and M. Azizoglu. Stochastic Modeling of TCP over Lossy Links. In *INFOCOM'00*, Tel-Aviv, Israel, March 2000.

[3] O. Ait-Hellal, E. Altman, D. Elouadghiri, M. Erramdani, and N. Mikou. Performance of TCP/IP : the case of two Controlled Sources. In *ICCC'97*, Cannes, France, November 1997.

[4] E. Altman, K. Avrachenkov, and C. Barakat. A stochastic model of TCP/IP with stationary ergodic random losses. Technical Report RR-3824, INRIA, November 1999.

[5] E. Altman, K. Avrachenkov, and C. Barakat. TCP in presence of bursty losses. Technical Report RR-3142, INRIA, July 1999.

[6] E. Altman, J. Bolot, P. Nain, D. Elouadghiri, P. B. M. Erramdani, and D. Collange. Performance Modeling of TCP/IP in Wide-Area Network. Technical Report RR-3142, INRIA, 1997.

[7] F. Baccelli and D. Hong. TCP is Max-Plus Linear. Technical Report RR-3986, INRIA, 2000.

[8] P. Brown. Resource sharing of TCP connections with different round trip times. In *INFOCOM'00*, Tel-Aviv, Israel, March 2000.

[9] N. Cardwell, S. Savage, and T. Anderson. Modeling TCP latency. In *INFOCOM'*, Tel-Aviv, Israel, March 2000.

[10] D. Comer. *Internetworking with TCP/IP, Volume 1 : Principles, Protocols, and Architecture, 3rd edition*. Prentice-Hall, 1995.

[11] S. Fortin and B. Sericola. A Markovian Model for the Stationary Behavior of TCP. Technical Report RR-4240, INRIA, September 2001. http://www.inria.fr/rrrt/rr-4240.html.

[12] P. Hurley, J. Y. L. Boudec, and P. Thiran. A Note on the Fairness of Additive Increase and Multiplicative Decrease. In *ITC-16*, Edinburgh, Scotland, June 1999.

[13] A. Kumar. Comparative Performance Analysis of Versions of TCP in a Local Networks with a Lossy Link. *IEEE/ACM Transactions on Networking*, 6(4), August 1998.

[14] 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*, 5(3), 1997.

[15] M. Mathis, J. Semke, J. Mahdavi, and T. Ott. The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm. *Computer Communications Review*, 27(3), July 1997.

[16] V. Misra, W. B. Gong, and D. Towsley. Stochastic Differential Equation Modeling and Analysis of TCP-Windowsize Behavior. In *Performance'99*, Istanbul, Turkey, October 1999.

[17] T. J. Ott, J. H. B. Kemperman, and M. Mathis. *The stationary behavior of ideal TCP congestion avoidance*, August 1996. http://www.argreenhouse.com/papers/tjo/.

[18] J. Padhye, V. Firoiu, and D. Towsley. A stochastic model of TCP Reno congestion avoidance and control. Technical Report 99-02, University of Massachussets, 1999.

[19] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose. Modeling TCP throughput : a simple model and its empirical validation. In *SIGCOMM'98*, Vancouver, Canada, September 1998.

[20] G. Rubino and B. Sericola. Sojourn times in Markov processes. *Journal of Applied Probability*, 26, 1989.

[21] W. R. Stevens. *TCP/IP Illustrated : Vol.1 The Protocols*. Addison-Wesley, 1994.

[22] W. R. Stevens. *TCP slow start, congestion avoidance, fast retransmit, and fast recovery algorithms*, January 1997. RFC 2001.

Proceedings of the 10th IEEE Int'l Symp. on Modeling, Analysis, & Simulation of Computer & Telecommunications Systems (MASCOTS'02)
1526-7539/02 $17.00 © 2002 **IEEE**