REFINED TCP PERFORMANCE EVALUATION WITH SIMPLE MODELING

Sophie FORTIN-PARISI and Bruno SERICOLA

IRISA-INRIA
Campus universitaire de Beaulieu
35042 Rennes cedex, France
Email : {Sophie.Fortin, Bruno.Sericola}@irisa.fr,

Abstract: This paper presents analytical results of a TCP (Transmission Control Protocol) model based on a Markov chain, refining the previous works on performance evaluation of one bulk transfer TCP flow among exogeneous traffic. While most of these works are mainly focused on the mean throughput evaluation, our model allows with low cost, a study of many other performance measures and thus a more detailed study of the TCP behavior.

Keywords: Performance modeling, Markov chain, Transmission Control Protocol, congestion control.

1 INTRODUCTION

The *Transmission Control Protocol* TCP represents a large part of today's Internet transfers. It has been, for that reason, the subject of many studies, centered either on live Internet measurements (downstream), simulations or modeling (upstream). TCP principle is to make sure that all data are actually received by the endpoint. When lost, a *segment* – i.e. a TCP packet – is retransmitted. Based on a sliding window dynamic, new segments are released into the network each time an acknowledgment (*ACK*), a small packet sent by the receiver to confirm the arrival of a segment, arrives. The function of TCP is to modify the window size, that can be correlated to an instantaneous throughput, according to an algorithm defined in the RFC2001 ([Stevens, 2001]) : an exponential increase (*slow start*) under a variable threshold, and then successive linear increases (*congestion avoidances*) separated by loss events that halve the window size.

A basic, but efficient, model presented in [Mathis et al, 1997] has shown that the mean throughput ρ of a TCP connection was in the order of $1/\sqrt{p}$, where p denotes the segment loss rate. Among earlier papers proposing a TCP model, many use a continuous-time and fluid approach ([Lakshman and Madhow, 1997], [Kumar, 1998], [Misra et al, 1999], [Altman et al, 1997], [Abouzeid et al, 1999] and [Altman et al, 1999]) and are usually and mainly interested in getting an analytical expression for the mean throughput of a single steady-state TCP connection. The case of multiple TCP connections is the subject of [Ait-Hellal et al, 1997], and [Brown, 2000] for instance, and an original modeling approach is provided in [Baccelli and Hong, 2000] by using the max-plus algebra.

Our paper is based on the reference works of [Padhye et al, 1998], [Padhye et al, 1999], and [Cardwell et al, 2000] which consider a discrete-time model and a discrete evolution of the window size. We present here the results of a discrete-time Markov chain model that we introduced in [Fortin and Sericola, 2001], and which aims to give analytical expressions not only for the mean throughput, but also for various performance measures, of a 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 server to another for instance).

The organization of this paper is as follows : after a presentation of the model in Section 2, we comment, in Section 3, our results for the mean throughput with a comparison to [Mathis et al, 1997] and [Padhye et al, 1998]. We then give in Section 4 other examples of performance measures which are the proportion of time during which the throughput is maximum, and the time-interval between two consecutive losses.

2 TCP MODELING

The choice of a discrete-time Markov chain modeling the congestion window evolution has been inspired from the pioneering work [Padhye et al, 1998], where the authors introduced the notion of *round* also used in [Padhye et al, 1999], [Cardwell et al, 2000] and [Fortin and Sericola, 2001]. A *round* is the period of time between the departure of the first segment of the current window and the arrival of its ACK. This defi-

Proceedings 17th European Simulation Multiconference
(c) SCS Europe BVBA, 2003 ISBN 3-936150-25-7

nition is coherent when the dispatch duration of all the segments and all the ACKs held in a given window is negligible compared to the *round trip time* RTT. Note that the duration of a round is thus close to the round trip time.

2.1 Presentation Of The Markov Chain

We model the window behavior by a discrete-time Markov chain $X = (X_n)_{n\geq 1}$ with two components $X_n = (W_n^c, W_n^{th})$. W_n^c denotes, when positive, the n-th round congestion window size and the null value for W_n^c is used to represent the time-out period. W_n^{th} denotes the value of the slow start threshold during the n-th round. We denote by b the number of segments validated per ACK. Typically, b is equal to 1 or 2 (in the case of delayed ACKs). TCP-Reno congestion control mechanisms can be described as follows:

- *slow start* (*ss*) : increase of 1 segment per ACK, that is $W_{n+1}^c = W_n^c + \lceil W_n^c/b \rceil$, as long as $1 \leq W_n^c < W_n^{th}$ and no loss occurs,
- *congestion avoidance* (*ca*) : increase of $1/W^c$ segment per ACK, that is increase of 1 segment every b rounds, as long as $W_n^{th} \leq W_n^c \leq W_{\max}$ and no loss occurs (when W_n^c reaches the maximum receiver's buffer capacity $W_{\max}$, it remains constant),
- segment loss detection by three duplicate ACKs (TD) : after the first ACK indicating that segment number n is the next expected one, the reception of three successive ACKs indicating that it is still missing notifies the loss of segment number n. $W_{n+1}^c = \max(\lfloor W_n^c/2 \rfloor; 1)$, $W_{n+1}^{th} = \max(\lfloor W_n^c/2 \rfloor; 2)$, and then a new congestion avoidance phase initiates,
- segment loss detection by *time-out* (TO) : when its ACK has not arrived before a timer (T_0) expiry : $W_{n+1}^c = 0$ and $W_{n+1}^{th} = \max(\lfloor W_n^c/2 \rfloor; 2)$, then enter a new *time-out period*,
- *time-out period* (*to*) : just after a TO detection, the segment is retransmitted as long as no ACK for this segment arrives (the timer value doubling from T_0 to $2T_0$, $4T_0$, $8T_0$, ... until $64T_0$), and then a new slow start phase begins with $W_{n+1}^c = 1$.

An illustration of the window evolution is given in Figure 1.

Because the state space E of the Markov chain is such that

$$E \subseteq \{0, \ldots, W_{\max}\} \times \{2, \ldots, W_{\max}/2\},$$

its size is less than or equal to $(W_{\max}+1)(W_{\max}/2-1)$. The number of states of E is thus less than or equal to 20000 for $W_{\max} < 200$, and less than or equal to 5000 for $W_{\max} < 100$. In both cases, the state space is quite small for the Markov chains computing methods.

2.2 Some Transition Probabilities

All transition probabilities of this Markov chain can be found in [Fortin and Sericola, 2001]. However, for a better understanding of this model, it is interesting to detail the two following phases. We denote by $P_{(i,j)(i',j')}$ the transition probability from state (i,j) to state (i',j').

2.2.1 *Time-out Period*

The time-out period corresponds to the case where $W_n^c = 0$. In order to make the mean duration of a time-out period equal to RTT times the mean number of successive visits to the state $(0,j)$, we define the two following transitions from each state $(0,j)$,

- $P_{(0,j)(0,j)} = 1 - \dfrac{RTT}{E[T_{to}]}$: lost segment not yet ACKed,
- $P_{(0,j)(1,j)} = \dfrac{RTT}{E[T_{to}]}$: the ACK has just arrived,

where

$$E[T_{to}] = T_0 \frac{f(p)}{1-p} - RTT$$

is the mean duration of a time-out period (see [Fortin and Sericola, 2001]), and

$$f(p) = 1 + p + p^2 + 4p^3 + 8p^4 + 16p^5 + 32p^6.$$

2.2.2 *Congestion Avoidance*

In congestion avoidance the congestion window is increased by 1 every b rounds, thus for a completely accurate model, we would have to define our model using three components : W_n^c, W_n^{th}, and a counter R_n going from 1 to b (see [Padhye et al, 1999]), with :

- $W_{n+1}^c = W_n^c$ and $R_{n+1} = R_n + 1$ if $R_n < b$ and no loss occurs (case 1),
- $W_{n+1}^c = W_n^c + 1$ and $R_{n+1} = 1$ if $R_n = b$ and no loss occurs (case 2).

However, that would first of all significantly increase the size of the Markov chain and thus any computing time. Secondly, that would not change the measures of interest since the stationary distribution on the state space of the original Markov chain remains the same if we define the transition probabilities such that the mean sojourn time of the Markov chain in a state (i,j), with $j \leq i < W_{\max}$, remains equal to b, that is, lasts b rounds. We thus have

- $P_{(i,j)(i,j)} = (1-p)^i \left(1 - \dfrac{1}{b}\right)$ (no loss, case 1),
- $P_{(i,j)(i+1,j)} = (1-p)^i \dfrac{1}{b}$ (no loss, case 2),

- $P_{(i,j)(0,max(\lfloor i/2 \rfloor,2))} = \left(1-(1-p)^i\right) q_i$ (TO-type loss),
- $P_{(i,j)(max(\lfloor i/2 \rfloor,1),max(\lfloor i/2 \rfloor,2))} = \left(1-(1-p)^i\right) q_i$ (TD-type loss),

where the probability that a loss (in a round of size i) is a TO-type loss is

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

if $i \geq 2b+2$, and $q_i = 1$ otherwise (see [Padhye et al, 1998]). This formula is obtained by the study of partial rounds, called the *residual rounds*.

This notion is based on the asumption that, when a segment loss occurs, all the following segments in its round get also lost, because the congestion responsible of that loss has not yet disappeared when the last segment of the round arrives. In Figure 2, if the $(k+1)$-th segment (and thus all the following ones) of the current window is lost, the k first segments will generate ACKs, and thus the congestion window will slide a little and release k new segments that form the residual round.

2.3 Stationary Distribution

Long term TCP transfers are supposed to reach a stationary regime. We will therefore focus on the cyclic stationary behavior of TCP (one *ss* phase, followed by successive *ca* phases until the next TO loss that causes a time-out period, and so on).

Note that, because of the exponential growth during slow start, the Markov chain does not reach all couples (i,j) for $0 \leq i \leq W_{\max}$ and $2 \leq j \leq \lfloor W_{\max}/2 \rfloor$. For instance, for $b = 1$ then the successive congestion window values in slow start are 1, $1+\lceil 1/b \rceil = 2$, $2+\lceil 2/b \rceil = 4$, 8, 16, 32, 64, ..., and for $b = 2$ they are 1, $1+\lceil 1/b \rceil = 2$, $2+\lceil 2/b \rceil = 3$, 5, 8, 12, 18, ... Excluding the states (i,j) which are not reached by the Markov chain, we obtain an irreducible and aperiodic finite state Markov chain. Therefore, the stationary probability distribution, denoted by π, exists and satisfies $\pi P = \pi$, where P is the transition probabilities matrix.

2.4 Results

Several measures of interest as, for instance, the speed of convergence to stationary regime, the proportion of time spent in slow start, the mean time-interval between two consecutive losses, the mean number of segments sent and received (successfully transmitted) between two losses or two time-out periods, the proportion of time in which the maximum window size is reached, and of course the mean throughput, can be expressed as functions of p, RTT, T_0 and the stationary probabilities $\pi(i,j)$. Some of them have been explored in [Fortin and Sericola, 2001]. We consider, in the following sections, the evaluation of the throughput, the mean time-interval between consecutive losses and the maximum window size.

3 THROUGHPUT COMPUTATION

3.1 Send Rate And Goodput

First of all, let us make an important distinction between the throughput in terms of number of segments sent per second which is called the *send rate* (the input rate) and denoted by ρ, and the throughput in terms of number of segments received by the endpoint which is called the *goodput* (the output rate) and denoted by ρ_0.

The send rate is given by the following formula

$$\rho = \frac{E[d_{to}] + E[d_{cycle}] + N_{loss}E[d_{rr}]}{E[T_{to}] + E[T_{cycle}] + RTT(N_{loss}-1)p_{rr}},$$

where :

- $E[d_{to}]$, $E[d_{cycle}]$ and $E[d_{rr}]$ denote the average number of segments sent during, respectively, each time-out period (*to*), each *cycle* (one *ss* and successive *ca* until the next TO-loss detection), and each residual round (*rr*),
- $E[T_{to}]$ and $E[T_{cycle}]$ denote the average duration of, respectively, each time-out period and each cycle,
- N_{loss} denotes the average number of losses per cycle,
- p_{rr} denotes the probability that a residual round is not empty, which means that at least one segment of the round that has experienced a loss, has been ACKed (the last residual round of a cycle, i.e. the one due to a TO-type loss, is not taken into account because it is considered as included in the following time-out period).

Similarly, the goodput is given by

$$\rho_0 = \frac{E[d^0_{cycle}] + N_{loss}E[d^0_{rr}]}{E[T_{to}] + E[T_{cycle}] + RTT(N_{loss}-1)p_{rr}},$$

where $E[d^0_{cycle}]$ and $E[d^0_{rr}]$ represent the mean number of segments successfully transmitted, respectively, during a cycle and during a residual round.

The expressions of all these quantities are detailed in [Fortin and Sericola, 2001].

For illustration, the expressions of the mean number of segments, respectively sent and received, during a cycle (between two successive time-out periods) are given by :

$$E[d_{cycle}] = \frac{\sum_{i=1}^{W_{\max}} i \sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} \pi(i,j)}{p_0 \sum_{j=2}^{\lfloor W_{\max}/2 \rfloor} \pi(0,j)},$$

and

$$E[d^0_{cycle}] = \frac{\frac{(1-p)}{p}\sum_{i=1}^{W_{\max}}(1-(1-p)^i)\sum_{j=2}^{\lfloor W_{\max}/2\rfloor}\pi(i,j)}{p_0\sum_{j=2}^{\lfloor W_{\max}/2\rfloor}\pi(0,j)},$$

where $p_0 = RTT/E[T_{to}] = P_{(0,j)(1,j)}$ (which means that $p_0\sum_{j=2}^{\lfloor W_{\max}/2\rfloor}\pi(0,j)$ is the probability that a cycle starts with the slow start threshold equal to j).

3.2 Comparison To Reference Models

Figure 3 shows that the results of our model are very close to the reference models presented in [Mathis et al, 1997] and [Padhye et al, 1998].

However, our results are slightly lower than theirs. This is explained by the accuracy of our model which, for instance, includes slow start phases and window size limitation. This difference is more obvious for lower RTT values, as shown in Figure 4.

The goodput gives similar results.

3.3 Efficiency

We call efficiency, the ratio $e = \rho_0/\rho$ (output rate over input rate). This ratio represents the percentage of useful data among the transfer load. The remaining load constitutes the retransmission of lost segments. Figure 5 shows the efficiency e for different values of $W_{\max}$. It confirms that, the higher the throughput is allowed to be (large $W_{\max}$), the more the transfer suffers losses.

4 OTHER EXAMPLES OF PERFORMANCE MEASURES

As we said in Section 2.4, many performance measures can be done with this model. Here we choose to present, in a first Section, the proportion of time $p_{\max}$ during which the congestion window size is maximum (the instantaneous send rate is $W_{\max}$ segments per RTT), and in a second Section, the time-interval between two consecutive losses.

4.1 Maximum Window Size

Figure 6 shows the evolution of

$$p_{\max} = \sum_{j=2}^{\lfloor W_{\max}/2\rfloor}\pi(W_{\max},j),$$

for different values of $W_{\max}$. Although it is not surprising that $p_{\max}$ is sensitive to $W_{\max}$, this figure shows that for high values of $W_{\max}$ and low values of p, neglecting a maximum size for the congestion window would not have much impact on the results. This is absolutely wrong for lower values of W_{max} and higher values of p, e.g. for $W_{\max} = 32$ and $p = 0.001$ we have $p_{\max} \simeq 33\%$.

This means that during one third of the time, the window size is equal to $W_{\max}$ and is not growing anymore. Any model that does not consider a window limitation will thus significantly overestimate the connection throughput.

4.2 Time-interval Between Two Consecutive Losses

Figure 7 shows the mean time-interval between two consecutive losses in a cycle, denoted by $E[\Delta T_{loss}]$, and equal to the mean duration $E[T_{ca}]$ of a congestion avoidance phase. In [Fortin and Sericola, 2001], we proved that

$$E[T_{ca}] = \frac{\sum_{j=2}^{\lfloor W_{\max}/2\rfloor}\Big(X_j(\alpha_{2j}+\alpha_{2j+1}) + X_{w_{n_j+1}}\beta_j\Big)}{\sum_{j=2}^{\lfloor W_{\max}/2\rfloor}(\alpha_{2j}+\alpha_{2j+1}+\beta_j)},$$

where

- $\alpha_i = \left(1-(1-p)^i\right)(1-q_i)\sum_{j=2}^{\lfloor W_{\max}/2\rfloor}\pi(i,j)$,
- $\beta_j = \pi(w_{n_j},j)(1-p)^{w_{n_j}}$,
- $X_i = RTT(1-p)^{-\frac{bi(i-1)}{2}}\left(\sum_{w=i}^{W_{\max}-1}\lambda_w + \mu\right)$,
- $\lambda_w = (1-p)^{\frac{bw(w-1)}{2}}\frac{1-(1-p)^{bw}}{1-(1-p)^w}$,
- $\mu = \frac{(1-p)^{\frac{bW_{\max}(W_{\max}-1)}{2}}}{1-(1-p)^{W_{\max}}}$.

When $W_{\max}$ increases, the rounds are likely to reach bigger sizes, and therefore, the risk of a segment loss also increases. That is why the bigger the $W_{\max}$, the higher the loss frequency, and the lower the $E[\Delta T_{loss}]$.

5 CONCLUSION

This paper is based on a Markov model, and extends the well-known discrete model of [Padhye et al, 1998] which is a reference in modeling the TCP stationary behavior. We have shown that our results for the mean throughput are consistent with previous works led on the subject.

However, we believe that we got more various and accurate results than many other models, without using neither too complex mathematical theories, nor too heavy computation methods. The examples of performance measures that we developed in this paper only represent an sample of what our model can bring. What is more, its strength also lies in its easy adaptability to other additive increase and multiplicative decrease parameters than $1/b$ and $1/2$, and also to other functions of increase and decrease with relatively reasonable modifications. Such a generalization will be the object of further work.

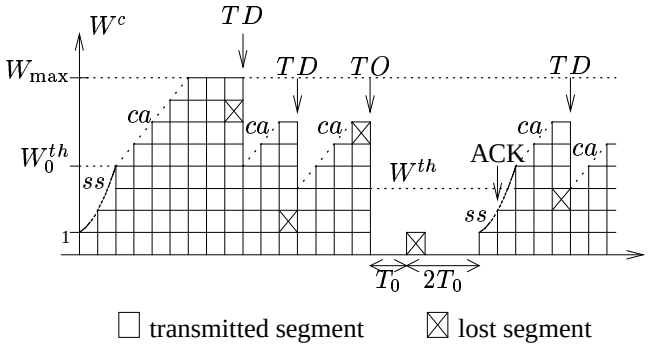

Figure 1: Example of congestion window evolution.

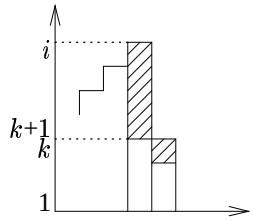

Figure 2: Residual round due to the ACKment of k segments.

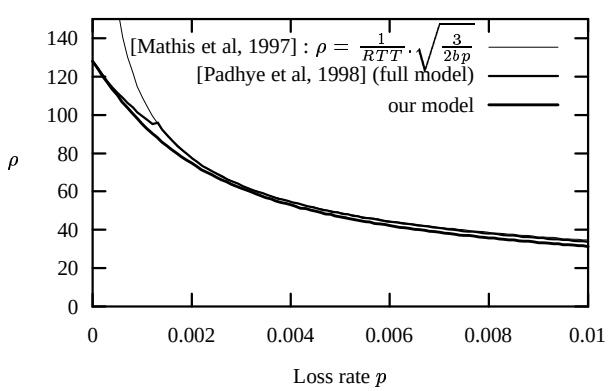

Figure 3: Send rate ρ vs previous models for $RTT = 0.250$ s ($W_{max} = 32$, $b = 2$, $T_0 = 0.500$ s)

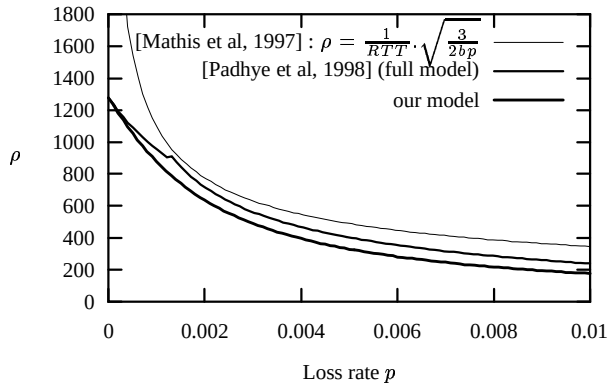

Figure 4: Send rate ρ vs previous models for $RTT = 0.025$ s ($W_{max} = 32$, $b = 2$, $T_0 = 0.500$ s)

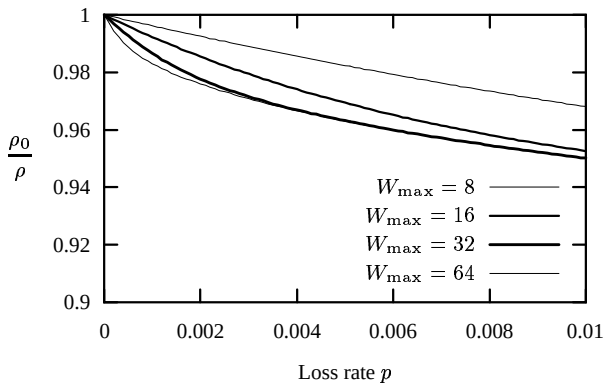

Figure 5: Efficiency $e = \rho_0/\rho$ for different values of $W_{\max}$ ($b = 2$, $RTT = 0.250$ s, $T_0 = 0.500$ s)

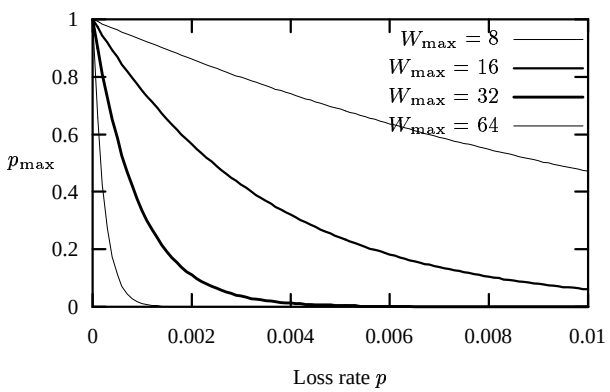

Figure 6: Evolution of $p_{\max}$ for different values of $W_{\max}$ ($b = 2$, $RTT = 0.250$ s, $T_0 = 0.500$ s)

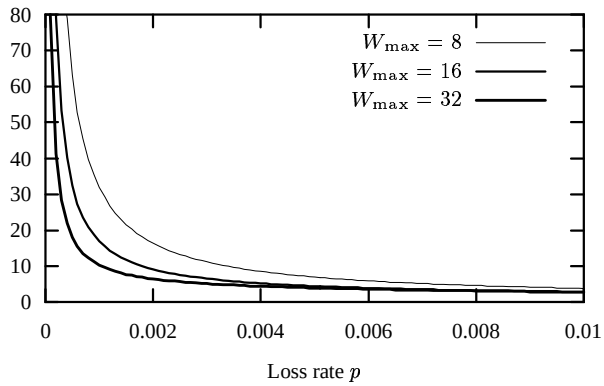

Figure 7: Mean loss interval $E[\Delta T_{loss}]$ in function of loss rate ($b = 2$, $RTT = 0.250$ s, $T_0 = 0.500$ s)

REFERENCES

Fortin S. and Sericola B. 2001, "A Markovian Model for the Stationary Behavior of TCP". *INRIA* RR-4240, *http://www.inria.fr/rrrt/rr-4240.html* .

Padhye J., Firoiu V., Towsley D. and Kurose J. 1998, "Modeling TCP Throughput : a simple model and its empirical validation". *In Proc. SIGCOMM'98* (Vancouver, Canada).

Padhye J., Firoiu V. and Towsley D. 1999, "A stochastic model of TCP Reno congestion avoidance and control". *University of Massachussets* **99-02**.

Cardwell N., Savage S. and Anderson T. 2000, "Modeling TCP latency". *In Proc. INFOCOM'00* (Tel-Aviv, Israel).

Stevens W. 1997, "TCP slow start, congestion avoidance, fast retransmit, and fast recovery algorithms". RFC 2001.

Lakshman T. V. and Madhow U. 1997, "The Performance of TCP/IP for Networks with High Bandwidth-Delay Products and Random Loss". *IEEE/ACM Transactions on Networking* **5**(3).

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

Baccelli F. and Hong D. 2000, "TCP is Max-Plus Linear". *INRIA* RR-3986.

Brown P. 2000, "Resource sharing of TCP connections with different round trip times". *In Proc. INFOCOM'00* (Tel-Aviv, Israel).

Altman E., Bolot J., Nain P., Elouadghiri D., Erramdani M., Brown P. and Collange D. 1997, "Performance Modeling of TCP/IP in Wide-Area Network". *INRIA* RR-3142.

Altman E., Avrachenkov K. and Barakat C. 1999, "TCP in presence of bursty losses". *INRIA* RR-3142.

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

Misra V., Gong W.-B. and Towsley D. 1999, "Stochastic Differential Equation Modeling and Analysis of TCP-Windowsize Behavior". *In Proc. Performance'99* (Istanbul, Turkey).

Abouzeid A. A., Roy S. and Azizoglu M. 2000, "Stochastic Modeling of TCP over Lossy Links". *In Proc. INFOCOM'00* (Tel-Aviv, Israel).

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

BIOGRAPHY

Sophie FORTIN−PARISI is teaching applied mathematics in the Telecommunications and Networks department of the Institute of Technology (IUT) of Valence (France) since 1998, and is preparing a Ph.D. supervised by Bruno SERICOLA. Her main research activity is in Internet flow control performance evaluation with stochastic models.

Bruno SERICOLA received the Ph.D. degree in computer science from the University of Rennes I in 1988. He has been with INRIA (Institut National de Recherche en Informatique et Automatique, a public research French laboratory) since 1989. His main research activity is in computer and communication systems performance evaluation, dependability and performability analysis of fault-tolerant architectures and applied stochastic processes.