TCP is Max-Plus Linear
and what it tells us on its throughput

François Baccelli* & Dohy Hong†
INRIA–ENS
ENS, DI, 45 rue d'Ulm,
75230 Paris Cedex 05, France.
[Francois.Baccelli,Dohy.Hong]@ens.fr

ABSTRACT
We give a representation of the packet-level dynamical behavior of the Reno and Tahoe variants of TCP over a single end-to-end connection. This representation allows one to consider the case when the connection involves a network made of several, possibly heterogeneous, deterministic or random routers in series. It is shown that the key features of the protocol and of the network can be expressed via a linear dynamical system in the so called max-plus algebra. This opens new ways of both analytical evaluation and fast simulation based on products of matrices in this algebra. This also leads to closed form formulas for the throughput allowed by TCP under natural assumptions on the behavior of the routers and on the detection of losses and timeouts; these new formulas are shown to refine those obtained from earlier models which either assume that the network could be reduced to a single bottleneck router and/or approximate the packets by a fluid.

1. INTRODUCTION
Various approaches have been investigated to characterize the key properties of the adaptive, additive increase, multiplicative decrease (AIMD) window flow control of TCP, including heuristics and simulations, fluid approximations or Markovian analysis [10, 11, 1, 13, 12, 8]. All analytical models are based on the so called *single bottleneck heuristic* [9]. It was also recently shown that window flow control on networks consisting of several routers in series admits a simple max-plus linear representation when window size is *constant* [2]. The present paper focuses on a class of models which combine the AIMD adaptive window size mechanism of TCP and a network model made of several routers in series. We show that their dynamics can be described at packet level via matrix recurrences in the max-plus algebra. Both the deterministic packet transmission time case and various stochastic models that have been used in the literature are considered, including the case where there are random losses in addition to losses due to buffer overflow, and the case when the packet transmission times are randomly perturbed by the rest of the traffic. The key aspects of the protocol can be represented, including congestion losses, timeouts, random losses, propagation and queueing delays as well as delays due to the flow control mechanism, window adaptation etc. We show how this approach allows one to establish general links between spectral properties of max-plus matrices and the mean throughput of TCP. This is used to derive closed form formulas for the maximal achievable throughput. These formulas can be used to analyze the case with several bottleneck routers. They are shown to be asymptotically compatible with the classical ones when the maximal window size tends to ∞. This framework allows one to analyze the instantaneous, possibly random fluctuations of the throughput, which may be useful for estimating the QoS offered to the connection. This approach is also shown to be particularly well suited for an efficient though detailed exact simulation of the end-to-end dynamics of the TCP protocol over large networks. In particular, it is proved that when using this approach, the simulation of n packets over K routers can be made with a computational cost of at most $2n(Kw^*)^2$, where w^* denotes the maximal window size.

In addition to these theoretical contributions, several phenomena of practical interest are also pointed out: (a) As soon as there are random perturbations due to cross traffic in the routers, *the throughput cannot be expressed in terms of the mean bitrate of the bottleneck router and the mean RTT only*: for example, permuting two routers along the route may then lead to a different throughput. (b) *A given overall loss probability is in general not enough to predict throughput*; in particular losses due to random perturbations created by cross traffic have a more severe effect on throughput than that of congestion losses stemming from a high send rate. (c) *Variance may have a significant effect on throughput*: keeping all mean service times fixed in the routers, an increase of variance may lead to a degradation of throughput.

The paper is structured as follows. In Section 2, we introduce the model and we give its representation in terms of a

*The work of this author was supported by the European Union TMR Alapedes project, RB-FMRX-CT-96-0074
†The work of this author was supported by a PhD grant from Ecole Polytechnique

linear max-plus recurrence equation. We then establish the main theoretical results of the paper by showing the link between TCP throughput and max-plus Lyapunov exponents. In Section 3, we consider the class of deterministic models and show periodicity results together with links between throughput and max-plus matrix eigenvalues. In Section 4, we consider two classes of stochastic models representing the random perturbations created by cross traffic. In Section 5, we give a brief list of further questions and extensions that can be treated along the same lines and for which analytical formulas extending those of the basic cases (or at least new simulation methods based on products of random matrices) can be expected.

2. MAX-PLUS REPRESENTATION

2.1 The max-plus algebra

Roughly speaking, the scalar max-plus "algebra" is the semi-ring structure over the real line where one replaces the addition by max (denoted $\oplus$) and the multiplication by plus (denoted $\otimes$). It is the fact that $\otimes$ is distributive w.r.t. $\oplus$ which allows one to extend classical concepts of linear algebra to this framework, and in particular matrix theory. This scalar semi-ring is denoted $(\mathbb{R}_{\max}, \oplus, \otimes)$, where $\mathbb{R}_{\max} = \mathbb{R} \cup \{-\infty\}$ is the real line completed by $-\infty$, the neutral element for $\oplus$. In what follows, we will denote $(\mathbb{R}_{\max}^{d,d}, \oplus, \otimes)$ the set of square matrices of dimension d in this algebra, where the two operations $\oplus$ and $\otimes$ have the following meaning when applied to matrices:

$$\begin{aligned} (A \oplus B)_{ij} &= A_{ij} \oplus B_{ij} = \max(A_{ij}, B_{ij}), \\ (A \otimes B)_{ij} &= \bigoplus_{1 \le k \le d} A_{ik} \otimes B_{kj} = \max_{1 \le k \le d} (A_{ik} + B_{kj}). \end{aligned}$$

For more details on this algebra, which is also used for QoS guarantees in networks [6], the reader may refer to [3] or [6].

2.2 The network model

Our basic model consists of a single source sending packets to a single destination over a path made of K routers in series. The transmission of the packets of this *reference flow* is assumed to be TCP controlled. Each router is represented by a single server queue. Each queue serves the packets of the reference flow as well as those of other flows, which will be referred to as *cross traffic* flows in what follows. Each router is assumed to be a FIFO queue for the packets of the reference flow. The nth packet of the reference flow arriving at queue i requires there an *aggregated service time* $\sigma_i(n)$. In case of a FIFO router, this aggregated service time captures both the processing time of this packet by the router and that of the backlog of cross traffic packets interleaved between the arrival time of packet # $n-1$ and that of packet # n in queue i. The model also incorporates some propagation delays between routers. The propagation delay from router i to j will be assumed to be deterministic and will be denoted $d_{i,j}$.

The input rate is controlled by a dynamic window size. This window mechanism controls the maximum number of packets sent by the source that have not been acknowledged by the destination.

2.3 From feedback to window

Let $ACK(n)$ denote the flow/congestion feedback signal giving information on the state of the network seen by packet # n. For example, $ACK(n) = 1$ if neither loss nor timeout are experienced by packet # n, otherwise $ACK(n) = 0$ which means either loss (LO) or timeout (TO).

In the deterministic case, at the time of the reception of signal $ACK(n)$ (either the reception of the acknowledgment of packet # n, or the detection of its loss or timeout), the window size is updated according to the following rule:

$$W(n) = F(W(n-1), ACK(n)), \quad n > 0, \tag{1}$$

with some initial condition $W(0) = 1$.

Once the sequence $\{ACK(n)\}$ is known, the above recurrence relation gives the *reference* window size process $\{W(n)\}$. Note that $W(n)$ is the *window size at the reception of the acknowledgment of packet # n* by the source. The associated *effective* window size is by definition

$$w_n = (\text{int})\ W(n) = \lfloor W(n) \rfloor. \tag{2}$$

We assume that the reference window size takes its values in a finite set with maximum element W^*. We will denote $w^* = \lfloor W^* \rfloor$.

We also assume that the evolution of the window size can be decomposed into two phases with the following properties:

increasing phase $(ACK = 1) : 0 \le F(W(n), 1) - W(n) \le 1$,
decreasing phase $(ACK = 0) : 1 \le F(W(n), 0) < W(n)$.

In the following, we allow F to depend on a sequence $\{\Theta(n)\}$, where $\Theta(n)$ gives the threshold that separates the slow-start phase from the congestion-avoidance phase. In this case, the pair $(W(n), \Theta(n))$ is updated according to the refined rule:

$$W(n) = F(\Theta(n-1), W(n-1), ACK(n)), \tag{3}$$
$$\Theta(n) = \phi(\Theta(n-1), W(n-1), ACK(n)), \tag{4}$$

Here are two ideal cases of particular interest, which represent simplified versions of Tahoe and Reno:

1. TCP Tahoe :

$$\begin{aligned} &\phi(\Theta(n-1), W(n-1), 1) = \Theta(n-1), \\ &\phi(\Theta(n-1), W(n-1), 0) = \lfloor \alpha W(n-1) \rfloor, \\ &F(\Theta(n-1), W(n-1), 1) = \\ &\qquad \begin{cases} \min(W(n-1)+1, W^*), & \text{if } W(n-1) < \Theta(n-1), \\ W(n-1), & \text{if } W(n-1) = W^*, \\ \min(W(n-1) + \frac{1}{w_{n-1}}, W^*), & \text{otherwise}. \end{cases} \\ &F(\Theta(n-1), W(n-1), 0) = 1. \end{aligned}$$

2. TCP Reno : the same as above but with the following adaptation:

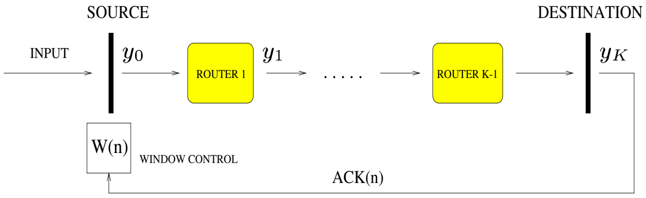
$$\begin{aligned} &F(\Theta(n-1), W(n-1), LO) = \max(1, \lfloor \alpha W(n-1) \rfloor), \\ &F(\Theta(n-1), W(n-1), TO) = 1. \end{aligned}$$

Here, $0 < \alpha < 1$ and $\Theta(0)$ are parameters and $W(0) = 1$. In the following examples, we will mainly consider the case $\alpha = 1/2$.

2.4 From window to dater

In what follows, we assume that the input queue is saturated (the non saturated case can be considered along the same lines as explained in §5). Then the network behaves as a closed network, the throughput of which gives the maximal rate at which the source can send packets while keeping the source buffer stable [2].

Let $x_i(n)$ be the time at which packet # n starts its aggregated service on router i (this is the time when this packet is head of the line within the set of packets of the reference flow).



Let $y_i(n) = x_i(n) + \sigma_i(n)$ be the time when packet # n leaves router i and let $\sigma_0(n) \equiv 0$. Let v_n be the *window size experienced by packet* # $n+1$, when it is sent by the source. In general, v_n and w_n (defined above) do not coincide. If the sequence $\{v_n\}$ is known, then $\{y_i(n)\}$, $0 \le i \le K$, $n \ge 1$, satisfies the equations :

$$y_0(n) = y_K(n - v_{n-1}) \otimes d_{K,0},$$
$$y_i(n) = [y_{i-1}(n) \otimes d_{i-1,i} \oplus y_i(n-1)] \otimes \sigma_i(n), \;\; i = 1, \dots, K.$$

In this model, the transmission of acks from the destination to the source is represented by a simple delay $d_{K,0}$. One can represent this backward route as a sequence of routers similar to that of the forward route with only slight modifications of the basic model.

We define $Y(n) = (y_1(n), y_2(n), .., y_K(n)) \in \mathbb{R}_{\max}^{1,K}$, and

$$Z(n) = (Y(n), Y(n-1), .., Y(n-w^*+1))^t \in \mathbb{R}_{\max}^{K.w^*,1}.$$

The last vector will be referred to as the *dater vector* in what follows.

Let M_i, $i \in \{1, .., w^*\}$, be given matrices of $\mathbb{R}_{\max}^{K,K}$. Below, $(M_1|M_2|\cdots|M_{w^*})$ denotes the block matrix of $\mathbb{R}_{\max}^{K.w^*,K.w^*}$, where blocks are of size $K \times K$; all blocks are equal to the matrix $\mathcal{E}$ of $\mathbb{R}_{\max}^{K,K}$ with all its entries equal to $-\infty$, but for the first line of blocks which is $M_1, M_2, \dots, M_{w^*}$.

LEMMA 1. [Max-plus representation] *If the system is initially empty, and if the sequence of experienced window sizes is $\{v_n\}_{n\in\mathbb{N}}$, then the dater vectors $Z(n)$ satisfy the following max-plus matrix recurrence relation:*

$$Z(n) = A_{v_{n-1}}(n) \otimes Z(n-1), \;\; n \ge 1, \tag{5}$$

where $Z(0) = (0, .., 0)^t$ and

$$\begin{aligned} A_1(n) &= (M(n) \oplus M'(n)|\mathcal{E}|..|\mathcal{E}) \oplus D, \\ A_2(n) &= (M(n)|M'(n)|\mathcal{E}|..|\mathcal{E}) \oplus D, \quad \dots, \\ A_{w^*}(n) &= (M(n)|\mathcal{E}|..|\mathcal{E}|M'(n)) \oplus D. \end{aligned}$$

In these formulas, $M(n)$ and $M'(n)$ are given by:

$$(M(n))_{ij} = \begin{cases} \sum_{k=j}^{i} \sigma_k(n) + \sum_{k=j}^{i-1} d_{k,k+1}, & \text{if } i \ge j, \\ -\infty, & \text{if } i < j, \end{cases}$$

$$(M'(n))_{ij} = \begin{cases} \sum_{k=1}^{i} (d_{k-1,k} + \sigma_k(n)) + d_{K,0}, & \text{if } j = K, \\ -\infty, & \text{if } j < K, \end{cases}$$

and D is the square matrix of dimension Kw^* with all its entries equal to $-\infty$ but for those of the form $D_{K+i,i}$, $i = 1, \dots, K(w^*-1)$, which are all equal to 0.

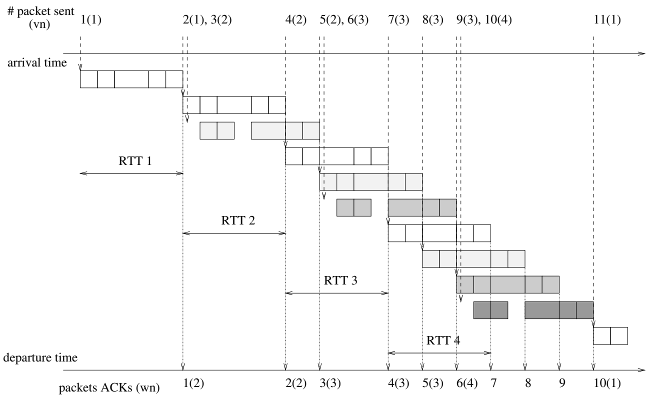
PROOF. The proof is immediate from the dynamics established for the $y_k(n)$ variables when expanding the max-plus product (5) coordinate by coordinate. □

REMARK 1. *At the level of representation adopted here, no difference exists between packets and retransmitted packets. In particular, we will make no difference between* send rate, throughput *or* goodput *[12].*

2.5 Example of evolution

Here is an explicit pathwise evolution of the dater vector and the window size: we take $K = 5$, $w^* = 4$ and $(\sigma_1(n), \cdots, \sigma_5(n)) = (1, 1, 2, 1, 1)$ for all n. We consider a periodic window size evolution (which is that of TCP Tahoe without slow start, cf. §3):

$$(v_0, v_1, v_2, \dots) = (1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 1, 1, 2, 2, 2, \dots).$$



For instance, packet # 6 experiences a window size of $v_5 = 3$, a fact that we denote 6(3) on the figure: this means that the admission of packet # 6 takes place at the time when the acknowledgment of packet $6 - 3 = 3$ is received.

2.6 From dater and window to feedback

2.6.1 *Deterministic feedback*

The deterministic model for the detection of losses and timeouts is:

$$ACK(n) = G(v_{n-1}, \overline{Z}(n)). \tag{6}$$

Here $\overline{Z}(n)$ is the equivalence class of the dater vector (defined in Lemma 1) for the equivalence relation: $Z \sim Y$ if for all i, $Z_i = Y_i + c$ for some constant c, whereas v_{n-1} is the window experienced by packet # n. Here are a few basic examples:

- **Rate based loss detection** Assume one can deduce some estimate $1/\sigma^*(n)$ of the current bottleneck service rate in the network, and some estimate $\mathrm{S}(n)$ of the current round trip time, both from $(v_{n-1}, \overline{Z}(n))$; then it makes sense to state that one detects a (congestion) loss when the average send rate $\frac{v_{n-1}}{\mathrm{S}(n)}$ reaches the bottleneck rate, namely

$$\begin{aligned}(\mathbf{G1}) : ACK(n) = 0 \text{ (Tahoe)}, \quad LO \text{ (Reno)},\\ \text{if } \frac{v_{n-1}}{\mathrm{S}(n)} > \frac{1}{\sigma^*(n)}.\end{aligned}$$

- **Buffer overflow detection** If there is a maximal buffer capacity of β_i for the reference flow on router i, then it makes sense to state that

$$\begin{aligned}(\mathbf{G2}) : ACK(n) = 0 \ \text{ (Tahoe)}, \quad LO \text{ (Reno)},\\ \text{if } \exists 1 \le i \le K, \ y_{i-1}(n) + d_{i-1,i} < y_i(n - \beta_i).\end{aligned}$$

  Note that either $\beta_i < w^*$ in which case the values of $y_{i-1}(n) - y_i(n-\beta_i)$ can be retrieved from $\overline{Z}(n)$ indeed; or $\beta_i \ge w^*$, and then no loss can ever occur for the reference flow on router i.

- **Timeout detection** In this case, the function G also admits the value $RTO(n)$ of the timer for packet # n as an additional argument; this variable is built from moving averages of the preceding RTT's by a recurrence relation (see [14] and [15]).

$$\begin{aligned}(\mathbf{G3}) : ACK(n) = 0 \ \text{ (Tahoe)}, \quad TO \ \text{ (Reno)},\\ \text{if } \ y_K(n) - y_0(n) + d_{K,0} > RTO(n).\end{aligned}$$

  Since $y_0(n) = y_K(n - v_{n-1}) + d_{K,0}$, this condition can also be retrieved from $(v_{n-1}, \overline{Z}(n))$ at least in the case when $v_{n-1} < w^*$ (in case $v_{n-1} = w^*$, one more step of the dater history is in fact needed).

- **Large service times or RTT** Here is a model in the same spirit as (G2) or (G3) but somewhat simpler. In case of random service times, it makes sense to assume that a packet of the reference flow experiences loss and/or timeout in case of large enough aggregated service time on one of the routers, or in case of large enough sum of its aggregated service times:

$$(\mathbf{G4})\text{-Tahoe} : ACK(n) = 0, \ \text{if } \sigma(n) \in \mathcal{B}, \ 1 \text{ otherwise},$$

  where $\sigma(n) = (\sigma_1(n), \ldots, \sigma_K(n))$ and where $\mathcal{B}$ is a certain subset of $\mathbb{R}^K$ expressing one of the above properties (e.g. $\sum_i \sigma_i > X$ for timeout or $\sigma_i > Y_i$ for some i, for loss created by a large cross traffic on some router etc.) and

$$(\mathbf{G4})\text{-Reno} : ACK(n) = \begin{cases} LO, & \text{if } \sigma(n) \in \mathcal{C}, \\ TO, & \text{if } \sigma(n) \in \mathcal{D}, \\ 1, & \text{otherwise}, \end{cases}$$

  where $\mathcal{C}$ and $\mathcal{D}$ are subsets of $\mathbb{R}^K$ in the same vein as above.

REMARK 2. *In what follows, we will always assume that the detection of loss is instantaneous, namely that the effect of the loss of packet # n in terms of window size is applied from packet # $n+1$ on. This is of course an approximation in comparison to what happens effectively via the triple duplicate mechanism.*

2.6.2 *Stochastic feedback*

In this second and more general case, the feedback signals $\{ACK(n)\}$ are also function of some random perturbations represented by an i.i.d. $\{0,1\}$-valued random sequence $\{\xi(n)\}$. More precisely, (6) is replaced by

$$ACK(n) = \Gamma(v_{n-1}, \overline{Z}(n), \xi(n)), \tag{7}$$

with $\Gamma(v_{n-1}, \overline{Z}(n), 1) = G(v_{n-1}, \overline{Z}(n))$ for both Tahoe and Reno and, for Tahoe $\Gamma(v_{n-1}, \overline{Z}(n), 0) = 0$, whereas for Reno

$$\Gamma(v_{n-1}, \overline{Z}(n), 0) = \begin{cases} LO, & \text{if } G(v_{n-1}, \overline{Z}(n)) \text{ is } 1 \text{ or } LO; \\ TO, & \text{if } G(v_{n-1}, \overline{Z}(n)) = TO. \end{cases}$$

We denote p the probability that $\xi(1) = 0$. The case with $p = 0$ leads back to the deterministic scheme described above.

This stochastic model is to be compared to that of [13], where a global loss probability is used to capture both random packet losses and losses due to congestion. In contrast, in this refined stochastic model, these two mechanisms are separately described: random packet losses constitute an i.i.d. process independent of all other elements of the network and are captured by the sequence $\{\xi(n)\}$ (p is then the probability that a packet is lost due to random perturbations), whereas congestion based losses are captured by the G function.

Models with random timeouts in place of (or in addition to) random losses can be considered along the same lines.

2.7 Global dynamics and throughput

2.7.1 *Simplified dynamics*

For the sake of easy exposition, we will first describe the global dynamics when making the approximation that $w_n = v_n$ for all n. We will see later on how to correct this. Under this simplification, the overall dynamics is constructive: if one knows $W(n-1)$ and $Z(n)$, then one can compute $ACK(n)$ from either (6) or (7); this in turns allows one to define $W(n)$ and hence w_n using (1) and (2). Finally, the knowledge of $v_n = w_n$ and $Z(n)$ allows one to compute $Z(n+1)$ thanks to (5). We summarize this in the following theorem which refers to the family of max-plus matrices $A_i(n)$, $n \ge 0$, $1 \le i \le w^*$, defined in Lemma 1 and to the functions F and Γ defined above.

THEOREM 1. *Under the foregoing assumptions, the sequence of vectors $\{Z(n), W(n)\}$ satisfies the recurrence relation*

$$Z(n) = A_{\lfloor W(n-1) \rfloor}(n) \otimes Z(n-1), \tag{8}$$
$$W(n) = F\left(W(n-1), \Gamma(W(n-1), \overline{Z}(n), \xi(n))\right), \tag{9}$$

$n \ge 1$, with initial condition $Z(0) = (0,..,0)^t$ and $W(0) = 1$. In these equations $\{\xi(n)\}$ is an i.i.d. $\{0,1\}$-valued sequence representing random losses ($\xi(n) \equiv 1$ in case there are no such losses).

REMARK 3. *Equations (8) and (9) are given here in the simplest case where F is not a function of $\Theta(n-1)$ and where Γ is not a function of $RTO(n)$. In order to handle the general case, one should of course add the evolution equations* *for the variables $\Theta(n)$ and $RTO(n)$ to these two recurrence relations.*

The equations in Theorem 1 are the basis for the algebraic simulation scheme alluded to in the introduction. Since the matrices $A_{w_{n-1}}(n)$ are of dimension Kw^*, and since only matrix-vector products are required (in addition to the computation of the F and G functions, the cost of which is here neglected), one can simulate the controlled transmission of n packets through a network of K routers in $2n(Kw^*)^2$ operations on a single processor. This can be significantly reduced when using the fact that the matrices are in fact quite sparse.

2.7.2 Exact dynamics

In order to describe the exact dynamics (namely that where one does not make the simplification $v_n \equiv w_n$ anymore), one should keep track of the history of the reference window size defined in §2.3. Let $\mathcal{W}(n) = (W(n), \ldots, W(n-w^*+1))$ be this history, with the convention $W(k) = 1$ if $k \leq 1$. The *experienced window size* v_n is then obtained by picking the integer part of the appropriate coordinate of the $\mathcal{W}(n)$ vector. Here is the generic part of the procedure allowing one to select the appropriate coordinate:

$$
\begin{aligned}
&v = \lfloor W(n) \rfloor;\\
&\texttt{for}\ (k = 1; k < v; k\texttt{++})\\
&\qquad \texttt{if}\ (\lfloor W(n-k+1) \rfloor == \lfloor W(n-k) \rfloor + 1)\\
&\qquad\qquad v\texttt{--};
\end{aligned}
$$

This procedure, which is that to be used during the increasing phase of the reference window size process, stems from the observation that v_n is equal to $\lfloor W(n) \rfloor$ if the window size does not change during the transmission of packet # n, and that the discrepancy between $\lfloor W(n) \rfloor$ and v_n increases of one unit each time $\lfloor W(n-k) \rfloor$ jumps up. The procedure to be used within periods where the window size decreases depends on the version of the protocol. For instance, in the Tahoe case, one simply resets the $\mathcal{W}$ vector to $(1, \ldots, 1)$ each time the window decreases. Detailed examples are studied below.

Denote $v_n = a(\mathcal{W}(n))$ this mapping. Under the foregoing assumptions, the sequence of vectors $\{Z(n), \mathcal{W}(n)\}$ satisfies a recurrence relation of the form

$$Z(n) = A_{a(\mathcal{W}(n-1))}(n) \otimes Z(n-1), \tag{10}$$

$$\mathcal{W}(n) = \mathcal{F}\left(\mathcal{W}(n-1), \mathcal{G}(\mathcal{W}(n-1), \overline{Z}(n), \xi(n))\right), \tag{11}$$

$n \geq 1$, with initial condition $Z(0) = (0, .., 0)^t$, for mappings $\mathcal{F}$ and $\mathcal{G}$ which are mere extensions of the F and G mappings to the histories of the variables under consideration (e.g. $[\mathcal{F}(\mathcal{W}(n-1), \ldots)]_1 = [\mathcal{W}(n)]_1 = W(n) = F(W(n-1), \ldots)$ and for $1 < i \leq w^*$, $[\mathcal{F}(\mathcal{W}(n-1), \ldots)]_i = [\mathcal{W}(n)]_i = [\mathcal{W}(n-1)]_{i-1} = W(n-i+1)$).

In what follows, the default assumption will be that of simplified dynamics.

2.7.3 Throughput and Lyapunov exponents

The instantaneous throughput fluctuates forever due to the adaptation of the window and/or changes in the cross traffic in the routers. By definition, the *mean throughput* λ of the controlled connection is the *long term averaging of the instantaneous throughput*, namely the limit

$$\lambda = \lim_{n \to +\infty} \frac{n}{y_K(n)} = \lim_{n \to +\infty} \frac{n}{\sum_{p=1}^{n} (y_K(p) - y_K(p-1))}, \tag{12}$$

when it exists. This is a natural definition given the fact that our model makes no difference between send rate and goodput.

The max-plus *Lyapunov exponent* γ of the sequence of matrices is defined as:

$$\gamma = \lim_{n \to +\infty} \frac{\max_{1 \leq l,k \leq K.w^*} \left(A_{v_{n-1}}(n) \otimes \cdots \otimes A_{v_0}(1)\right)_{l,k}}{n}. \tag{13}$$

A sufficient condition for the the limit defining γ to exist (in the almost sure sense) is that the sequence $\{A_{v_{n-1}}(n)\}$ converges (with so called shift coupling) to some stationary and ergodic sequence. The existence then follows from Kingman's subadditive ergodic theorem (cf. [3]).

Since with our definition of $Z(0)$, $\max_{1 \leq l,k \leq K.w^*} \left(A_{v_{n-1}}(n) \otimes A_{v_{n-2}}(n-1) \otimes \cdots \otimes A_{v_0}(1)\right)_{l,k} = y_K(n)$, we see that under this coupling convergence property, the mean throughput is well defined and that it coincides with the inverse of the Lyapunov exponent of the sequence of matrices: $\lambda = \gamma^{-1}$.

It is beyond the scope of the present paper to give the minimal conditions for these convergences to hold, and we will rather analyze this question case by case.

3. DETERMINISTIC MODELS

In this section we assume that $\sigma_i(n) = \sigma_i$, for all n, where σ_i is non-random (this is the deterministic service time assumption) and that $\xi(n) \equiv 0$ (deterministic feedback assumption).

We will use the following notations: $\sigma^* = \max_{1 \leq i \leq K} \sigma_i$ and $\mathrm{S} = \sum_{i=1}^{K} \sigma_i$. For the sake of simple presentation, we will first consider the case when all propagation delays $d_{i,j}$ are 0, and then show how the formulas should be modified to cover the non-zero case.

THEOREM 2. *Assume that the service and the transmission times are rational numbers. Then under any of the above assumptions concerning the protocol (e.g. Reno or Tahoe with or without slow start), and the form of feedback (e.g. G2, or G2 and G3, or G1, etc.), the sequence of reference windows $\{w_n\}$ becomes ultimately periodic, with values in an integer interval of the form $[a^*, b^*]$, such that $1 \leq a^* \leq b^* \leq w^*$.*

The proof is forwarded to §7.

In what follows, we will assume that the period is made of a single increasing phase. This is always the case under (G1). The more complex periodic patterns which can take place under (G2) or (G3) can be studied in similar terms.

Denote t_i the number of occurrences of $a^* \leq i \leq b^*$ during a period and T the period: $T = t_{a^*} + t_{a^*+1} + \cdots + t_{b^*}$.

Thanks to Theorem 2, under the foregoing deterministic assumptions, one can simplify the dynamical system (8)-(9) by reducing it to the pure max-plus recurrence relation

$$Z(n) = A_{w_{n-1}}(n) \otimes Z(n-1), \tag{14}$$

where $\{w_n\}$ is the periodic window size in this theorem (see the examples below). Therefore, there exists a square matrix A' of dimension Kw^* describing the transient phase of the window size, and an integer m describing the number of packets sent in this transient phase, such that for all $n \geq 0$,

$$Z(nT+m) = \left(A_{b^*}^{t_{b^*}} \otimes A_{b^*-1}^{t_{b^*-1}} \otimes \cdots \otimes A_{a^*}^{t_{a^*}}\right)^n \otimes A' \otimes Z(0). \tag{15}$$

We then have the following theorem which establishes the link between the mean throughput of our deterministic TCP model and max-plus matrix eigenpairs (see [3] for more on the computation of eigenvalues and eigenvectors):

THEOREM 3. *If the square matrix (of dimension $K.w^*$) $A_{b^*}^{t_{b^*}} \otimes A_{b^*-1}^{t_{b^*-1}} \otimes \cdots \otimes A_{a^*}^{t_{a^*}}$ has a unique max-plus eigenvalue γ, then the mean throughput is $\lambda = \frac{T}{\gamma}$.*

3.1 Tahoe and Reno examples

3.1.1 TCP Tahoe

We first consider the TCP Tahoe model without the slow-start phase. Either $v_n = w^*$ for n large enough, or we have

$$\{v_1, v_2, \dots\} = \{1, 2, 2, 3, 3, 3, \dots, \underbrace{b^*-1, .., b^*-1}_{b^*-1 \text{ times}}, \underbrace{b^*, .., b^*}_{t_{b^*} \text{ times}}, 1, 2, 2, ..\}. \tag{16}$$

The value of b^* and that of t_{b^*} depend on the chosen feedback model (for instance, in the (G1) case, $t_{b^*} = 1$ and $b^* = \lfloor \frac{\mathrm{S}}{\sigma^*} \rfloor + 1$; see the proof of Theorem 2). In other words, $\forall i \in \{1, .., b^*-1\}$, $t_i = i$ and $1 \leq t_{b^*} \leq b^*$. Therefore $T = \frac{b^*(b^*-1)}{2} + t_{b^*}$.

COROLLARY 1. [Periodic TCP Tahoe without slow start] *Either the window is always equal to w^*, after a certain rank, in which case the mean throughput is*

$$\lambda = \min\left(\frac{1}{\sigma^*}, \frac{w^*}{\mathrm{S}}\right), \tag{17}$$

or there is an infinite number of epochs when the window drops to 1 and the mean throughput is:

$$\lambda = \frac{1}{2} \frac{b^*(b^*-1) + 2t_{b^*}}{\sum_{k=1}^{b^*-1} \max(\mathrm{S}, k\sigma^*) + t_{b^*}\sigma^*}. \tag{18}$$

A partial proof of Corollary 1 is given in §7 under (G1). The proof is based on the computation of the eigenvalue γ defined in Theorem 3, which is unique in this case. We also give a graphical interpretation of this eigenvalue property below.

REMARK 4. *Note that for (G1),*

$$\lambda = \frac{1}{2} \frac{b^*(b^*-1) + 2}{(b^*-1)\mathrm{S} + \sigma^*}, \tag{19}$$

so that λ only depends on σ^* and S (since $b^* = \lfloor \frac{\mathrm{S}}{\sigma^*} \rfloor + 1$). In this case, when $b^* \to +\infty$, the asymptotic throughput is such that $\lambda \sim \frac{1}{2}\frac{1}{\sigma^*}$. As for (G2) or (G3), there are no closed form expressions for b^*, which can nevertheless be computed numerically in $2K^2(w^*)^3$ operations. For these more complex models, λ depends in general on $\sigma = (\sigma_1, \dots, \sigma_K)$ and on $\beta = (\beta_1, \dots, \beta_K)$ as well as on the way $RTO(n)$ is updated.

3.1.1.1 Pathwise interpretation

In Figure 1 below (where we assume (G1) and $b^* = 6$), we restrict our attention to the pathwise evolution of the entrance time $y_0(n)$ and the departure time $y_K(n)$ of packet # n, which turn out to have more regularity than the daters associated with internal routers; the above eigenpair property receives the following interpretations: before congestion detection, packets sent behave as if there were no interactions between them, except for the pairs of packets sent at the same moment, i.e. when the window increases of one unit; for these pairs, the second packet always leaves the network σ^* units time later than the first one. Using this, one can read the eigenvalue property directly on the figure.

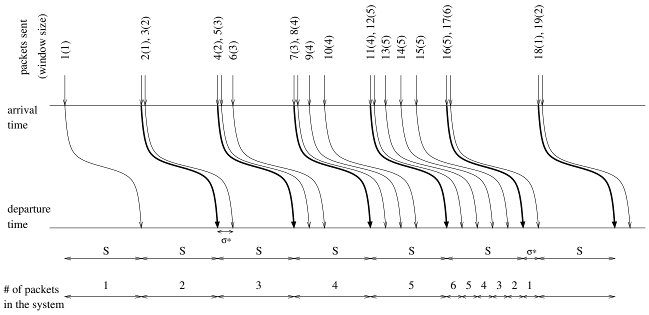


Figure 1: Interpretation through pathwise evolution

The sequence for Tahoe with slow start is:

$$\{1, 2, .., \theta-1, \underbrace{\theta, .., \theta}_{\theta \text{ times}}, \cdots, \underbrace{b^*-1, .., b^*-1}_{b^*-1 \text{ times}}, b^*\}, \tag{20}$$

with $\theta = \lfloor b^*/2 \rfloor$. The limitations of the simplified dynamics appear clearly here since this in fact leads to an "instantaneous" slow start phase where θ packets are sent at the same time.

COROLLARY 2. [Periodic TCP Tahoe with instantaneous slow start] *Under (G1), the throughput of TCP Tahoe with slow-start is given by:*

$$\lambda = \frac{1}{2} \frac{b^*(b^*-1) - \lfloor \frac{b^*}{2} \rfloor \left(\lfloor \frac{b^*}{2} \rfloor - 3\right)}{(b^*+1-\lfloor \frac{b^*}{2} \rfloor)\mathrm{S} + (\lfloor \frac{b^*}{2} \rfloor - 1)\sigma^*}. \tag{21}$$

When $b^* \to +\infty$, the asymptotic throughput is such that $\lambda \sim \frac{3}{4}\frac{1}{\sigma^*}$.

3.1.1.2 Exact dynamics

Under (G1) when moving from the simplified dynamics to the exact one, the window size that should be used in place of (16) is:

$$\{v_1, v_2 \dots\} = \{1, 1, 2, 2, 2, 3, 3, 3, 3, \dots, \underbrace{b^*-1, .., b^*-1}_{b^* \text{ times}}, b^*, 1, 2, 2, ..\}. \tag{22}$$

Then, the throughput is given by:

$$\lambda = \frac{1}{2}\frac{b^*(b^*+1)}{b^*\mathrm{S} + (b^*-1)\sigma^*}. \qquad (23)$$

For the same model with slow start, the following sequence should be used in place of (20):

$$\{1,1,2,2,..,\theta\text{-}1,\theta\text{-}1,\ \underbrace{\theta,..,\theta}_{\theta+1 \text{ times}},\ldots,\underbrace{b^*\text{-}1,..,b^*\text{-}1}_{b^* \text{ times}},b^*,1,1,2,2,..\}. \qquad (24)$$

One can check that these modifications do not change the asymptotic value of the throughput when letting $b^* \to \infty$.

Example 1. *Take TCP Tahoe without slow-start phase over 4 tandem queues with* $\sigma_1 = 3.2$, $\sigma_2 = 4.61$, $\sigma_3 = 2.7$, $\sigma_4 = 4.61$. $b^* = 4$, $w_n \in \{1,2,3,4\}$. *The throughput is equal to* 0.140084 *(Corollary 1). This is to be compared to the throughput given by* (23)*:* 0.134571.

3.1.2 TCP Reno

The periodic and deterministic evolutions of TCP Reno have been considered in [11] to get a heuristic value of the throughput. The above max-plus representation leads to a new formula that refines that of [11].

Corollary 3. [Periodic TCP Reno] *Under (G1) without slow start, the throughput of TCP Reno is given by:*

$$\lambda = \frac{1}{2}\frac{b^*(b^*-1) - \lfloor\frac{b^*}{2}\rfloor\left(\lfloor\frac{b^*}{2}\rfloor - 1\right) + 2}{(b^* - \lfloor\frac{b^*}{2}\rfloor)\mathrm{S} + \sigma^*}. \qquad (25)$$

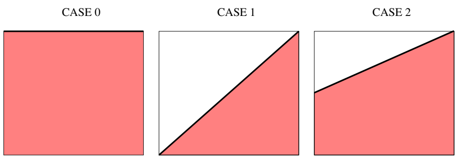
When $b^* \to +\infty$, *the asymptotic throughput is such that* $\lambda \sim \frac{3}{4}\frac{1}{\sigma^*}$. *In case of Reno with slow start, the formula is the same as that of Tahoe with slow start.*

3.2 Interpretation and comparison to earlier results

The results of this section are all under (G1).

3.2.1 Graphical interpretation of throughputs

The asymptotic throughputs found in Corollary 1 (case 1), Corollary 3 (case 2) and Corollary 2 (case 3) under (G1) have a natural graphical interpretation from a fluid approximation of the window size evolution: let $d_0 = \frac{1}{\sigma^*}$ be the throughput for static window size $w_n = b^*$ (case 0).



Graphical interpretation of throughputs

When w_n increases linearly from 1, the quantity of transmitted packets, which is proportional to the integral of $W(t)$ on a period, is indeed equal to $\frac{1}{2}d_0$ (case 1); when w_n increases linearly from $\frac{b^*}{2}$, the quantity of transmitted packets is scaled down by a factor $\frac{3}{4}$ (case 2).

3.2.2 Loss probability

The well known formula of the throughput for a single TCP connection in terms of loss probability p_{loss} and round trip time RTT is of the form [11]:

$$D_{th} = \frac{c_o}{RTT\sqrt{p_{loss}}},$$

where c_o is a real constant. For our deterministic model, we have: $RTT = \mathrm{S}$ and

$$\begin{aligned}
\text{case 1} &: \ p_{loss} = \frac{2}{b^*(b^*-1)+2},\\
\text{case 2} &: \ p_{loss} = \frac{2}{b^*(b^*-1) - \lfloor\frac{b^*}{2}\rfloor\left(\lfloor\frac{b^*}{2}\rfloor - 1\right) + 2},\\
\text{case 3} &: \ p_{loss} = \frac{2}{b^*(b^*-1) - \lfloor\frac{b^*}{2}\rfloor\left(\lfloor\frac{b^*}{2}\rfloor - 3\right)}.
\end{aligned}$$

When $b^* \to \infty$, we have $\sqrt{p_{loss}} \sim \frac{\sqrt{2}}{b^*}$ (case 1) and $\sqrt{p_{loss}} \sim \sqrt{\frac{8}{3}}\frac{1}{b^*}$ (cases 2 and 3). Therefore we have the following values for c_o:

case 1: $c_o = \frac{1}{\sqrt{2}} \simeq 0.71$; cases 2 and 3: $c_o = \sqrt{\frac{3}{2}} \simeq 1.22$.

Thus, for large values of b^* (or small value of p_{loss}), the asymptotic formula of Corollary 3 reduces to the formula in [11].

3.2.3 Extension to non zero propagation delays

All the results concerning (G1) hold with constant propagation delays $d_{i,j}$ when replacing the value of S by $\mathrm{S} = d_{K,0} + \sum_{k=1}^{K}(\sigma_k + d_{k-1,k})$.

3.2.4 Comparison with NS

The mean throughput obtained for these deterministic models can be compared to that given by the NS simulator when choosing an arbitrary packet size and when taking a bitrate for router i corresponding to σ_i. The send rates obtained from NS simulation and from our formulas may only differ due to discrepancies on the loss/congestion detection mechanism (discrepancies stem from the instantaneous loss detection assumption (see Remark 2) and also from the fact that we take the integer part of $W(n)$ rather than $W(n)$ etc.). However, for all deterministic models with the same periodic evolution of $\{v_n\}$, the evolutions are exactly the same. Here is an example: on NS we take a TCP connection with ftp source: $K = 10$, packet size is 1250 (40 for ack), buffer size is 2, all $d_{i,j}$ are equal to 0.1ms except for $d_{K,0}$ which is equal to 1ms; the bitrates are: $(10,5,4,2,5,4,5,5,4,5,5)$Mb for the links $0-1,..,9-10,10-0$. At $t = 100$s, NS gives 152.27 packets/s. For this example, $\mathrm{S} = 25.5$ms and $\sigma^* = 5$ms; using (G1), we get from (21): 134 packets/s. However, we note that b^* is actually equal to 7 in the NS simulation (since one RTT is needed to detect triple-acks) whereas it is equal to 6 in our model (this is precisely the difference between instantaneous and non instantaneous loss detections); taking (21) with $b^* = 7$ gives 152.55 packets/s.

4. STOCHASTIC MODELS

4.1 Deterministic services, random feedbacks

We now consider the case with simplified dynamics, with all service times still deterministic and rational, but with random feedback as defined in §2.6.

Under our assumptions, the sequence $\{(W(n), \overline{Z}(n))\}$ (resp. the sequence $\{(\Theta(n),W(n),RTO(n),\overline{Z}(n))\}$ when applicable) forms a Markov chain with finite state space Δ. If this Markov chain is irreducible, then the sequence of random matrices in (5) converges to a stationary and ergodic sequence in a sense which guarantees the existence of the mean throughput. More directly, if one denotes π the stationary probability of the Markov chain, then it follows from (12) that the inverse of the mean throughput can be expressed as

$$\lambda^{-1} = \gamma = \sum_{(w,\overline{z})\in\Delta} \alpha(\overline{z})\pi(w,\overline{z}), \tag{26}$$

where $\alpha(\overline{z}) = \overline{z}_K - \overline{z}_{2K}$ (see also [4]). Here is a concrete application of this general idea:

THEOREM 4. *Under (G1), if $p > 0$, $\{W(n)\}$ is an irreducible Markov chain on the integer interval $[1, b^*]$, with $b^* = \lfloor \frac{\mathrm{S}}{\sigma^*} \rfloor + 1$, and the throughput depends on service times only through* S *and σ^*.*

PROOF. The first property is immediate (the irreducibility stemming here from the fact that $\{W(n)\}$ can reach the value 1 from any initial condition by sufficiently many random losses in series). The last property follows from (26) and from the fact that the sequence $\{y_K(n+1) - y_K(n)\}_{n\in\mathbb{N}}$ takes its values in the set $\Phi = \{\mathrm{S} - (k-1)\sigma^*, k = 1, .., b^* - 1\} \cup \{\sigma^*\}$ (see Lemma 2 and the proof of Corollary 1 in the appendix). $\square$

4.1.1 Tahoe example

COROLLARY 4. [Markov TCP Tahoe without slow start] *Under (G1) the throughput of Tahoe without slow start is given by:*

$$\lambda = \frac{1}{\sigma^* + \sum_{k=1}^{b^*-1} [\mathrm{S} - k\sigma^*]\, q(k)},$$
$$\text{with} \quad q(k) = \frac{p(1-p)^{\frac{k(k+1)}{2}-1}}{1-(1-p)^{\frac{b^*(b^*-1)}{2}+1}}. \tag{27}$$

PROOF. In this case, $\{W(n)\}$ is an irreducible Markov chain on the set $X = \{1, 2, 2 + \frac{1}{2}, 3, 3 + \frac{1}{3}, 3 + \frac{2}{3}, 4, \ldots, b^* - 1 + \frac{b^*-2}{b^*-1}, b^*\}$. Let us denote by $\mu(x), x \in X$, the stationary probability of this Markov chain. Simple calculations give: for all $k + \frac{j}{k} \in X$ $(j = 0, \ldots, k-1;\ k = 2, \ldots, b^*)$,

$$\mu(k + \frac{j}{k}) = (1-p)^{\frac{(k-1)k}{2}+j}\mu(1),$$
$$\text{with} \quad \mu(1) = \frac{p}{1-(1-p)^{\frac{b^*(b^*-1)}{2}+1}}.$$

So we have

$$\gamma = \sum_{k=1}^{b^*-1} (\mathrm{S} - (k-1)\sigma^*)\mu(k + \tfrac{k-1}{k}) + \sigma^* \sum_{k=2}^{b^*}\sum_{j=1}^{k-1} \mu(k + \tfrac{j-1}{k}). \tag{28}$$

For $k = 1, \ldots, b^* - 1$, if we put $q(k) = \mu(k + \frac{k-1}{k})$, (28) immediately gives (27). $\square$

REMARK 5. *When $b^* \to \infty$, the asymptotic throughput takes a simple form if $p \sim \frac{2}{(b^*)^2}$; in this case,*

$$\lambda \sim \left(\frac{1-e^{-1}}{2\int_0^1 e^{-t^2}dt}\right)\frac{1}{\sigma^*} \simeq 0.42\frac{1}{\sigma^*}. \tag{29}$$

4.1.2 The impact of random losses

These Markov models can be used to show that the effect of losses due to random perturbations is preponderant compared to that of losses due to a too high send rate: indeed, the global loss probability of this model is: $\mu(b^*) + p(1 - \mu(b^*)) = \mu(1)$, where $\mu(b^*)$ is the loss due to congestion and $p(1 - \mu(b^*))$ is the loss due to random perturbations. For $\mu(1)$ fixed, Figure 2 shows how the throughput obtained by (27) decreases in p (this case is that when $\sigma^* = 1$, so that $\mathrm{S} = b^* - 1$).

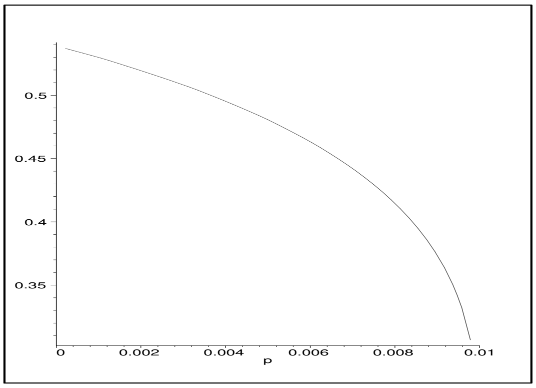


Figure 2: Throughput as a function of p for $\mu(1) = 0.01$

4.1.3 Reno examples

Similar results can be derived for TCP Reno type models. Due to the lack of space we will limit ourselves to a few numerical examples.

EXAMPLE 2. *Take $K = 4$ with $\sigma_1 = 3.2$, $\sigma_2 = 4.61$, $\sigma_3 = 2.7$, $\sigma_4 = 4.61$. $b^* = 4$, $w_n \in \{1, 2, 3, 4\}$. Figure 3 shows the evolution of $\frac{n}{y_4(n)}$ and w_n for Markov TCP Reno with $p = 0.1$.*

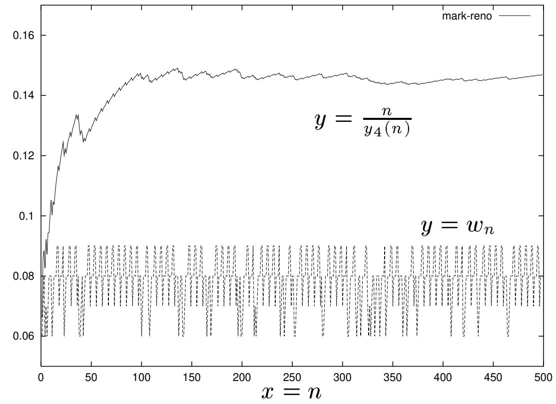


Figure 3: TCP Reno

4.2 Random service times

We now consider the case with random (aggregated) services on all routers. This case is the most difficult one, even for a constant window size. The difficulty stems in particular from the fact that the computational cost of the formulas grows in non-polynomial way with the maximum window size. We will assume here that the sequence $\{\sigma_i(n),\ i = 1,\dots,K\}_n$ is i.i.d. This is a simplified model w.r.t. our initial motivations where aggregated service times represent the influence of cross traffic on the packets of the reference connection (see [2]). Under this assumption, the sequence of matrices $\{A_i(n),\ i = 1,\dots,w^*\}_n$ is i.i.d. and for all loss detection models described in §2.6, $\{(W(n),\overline{Z}(n))\}$ (resp. the sequence $\{(\Theta(n),W(n),RTO(n),\overline{Z}(n))\}$ when applicable) forms a Markov chain.

4.2.1 Examples

4.2.1.1 Tahoe (G1)-(G4)

The assumptions concerning the feedbacks are (G1) and (G4);

- (G1) accounts for the losses due to an excessive send rate of the reference flow; we will take here $S(n) = \mathrm{S} = \mathbb{E}(\sum_{i=1}^{K}\sigma_i(1))$ and $\sigma^*(n) = \sigma^* = \max_{i=1}^{K}\mathbb{E}(\sigma_i(1))$, so that b^* is still given by $b^* = \lfloor \frac{\mathrm{S}}{\sigma^*} \rfloor + 1$;
- (G4) accounts for the losses and timeouts due to the variations of cross traffic.

For $n \le \frac{(b^*-1)b^*}{2} + 1$, let $\widetilde{B}(n)$ be the vector

$$\widetilde{B}(n) = \widehat{A}_{v_n}(n+1) \otimes \widetilde{A}_{v_{n-1}}(n) \otimes \cdots \otimes \widetilde{A}_{v_1}(2) \otimes Z(0),$$

where $v_1, v_2, \dots$, is the Tahoe sequence $(1,2,2,3,3,3,\dots,b^*-1,b^*-1,b^*)$. In this formula, the matrices $\{\widetilde{A}_i(n),\ i = 1,\dots,K\}$ (resp. $\{\widehat{A}_i(n),\ i = 1,\dots,K\}$) are i.i.d. and defined as $\{A_i,\ i = 1,\dots,K\}$, but when using the i.i.d. random variables $\widetilde{\sigma}(n)$ (resp. $\widehat{\sigma}(n)$) in place of $\sigma(n)$, where $\widetilde{\sigma}(n)$ (resp. $\widehat{\sigma}(n)$) is a random vector with the law of $\sigma(n)$ conditional on the property that $\sigma(n) \notin \mathcal{B}$ (resp. that $\sigma(n) \in \mathcal{B}$). For instance, if $\mathcal{B}$ is the set

$$\mathcal{B} = \{\sigma \in \mathbb{R}^K \text{ s.t. } \sigma_i > X \text{ for some } i\},$$

and if the random variables $\sigma_i(n)$, $i = 1,\dots,K$ are independent and uniformly distributed on the interval $[0,U]$, with $X < U$, then the random variables $\widetilde{\sigma}_i(n)$, $i = 1,\dots,K$ are still i.i.d. and uniform on the interval $[0,X]$, whereas the random variables $\widehat{\sigma}_i(n)$, $i = 1,\dots,K$ have a joint distribution which can be computed explicitly using order statistics.

Corollary 5. [TCP Tahoe with random service times] *Under (G1)-(G4), the throughput of TCP Tahoe without slow start is:*

$$\lambda = \frac{\sum_{k=1}^{b^*}\sum_{i=0}^{k-1} p_{ki}\left(\frac{(k-1)k}{2}+i+1\right)}{\sum_{k=1}^{b^*}\sum_{i=0}^{k-1} p_{ki}\ \mathbb{E}\left[\left(\widetilde{B}\left(\frac{(k-1)k}{2}+i+1\right)\right)_K\right]}, \tag{30}$$

where

$$p_{ki} = \pi^{\frac{(k-1)k}{2}+i}(1-\pi), \tag{31}$$

with $\pi = P(\sigma(n) \in \mathcal{B})$.

Proof. Since the matrix $M(n) \oplus M'(n)$ has the so called *memory loss property* on the set $\{(y_1,..,y_K) \in \mathbb{R}^K, y_1 \le .. \le y_K\}$ (see [4]), the stochastic process $\{W(n),\overline{Z}(n)\}$ is a regenerative process where the regeneration times are the epochs when the window size is equal to 1. The lengths T_l, $l \ge 1$, of the successive regeneration cycles are i.i.d. and such that

$$T_1 \in \left\{1,\dots,\frac{b^*(b^*+1)}{2}\right\}.$$

Let p_{ki} denote the probability that T_1 is equal to $\frac{k(k-1)}{2}+i+1$, $0 \le i < k \le b^*$. One obtains the value given in (31) for p_{ki} when using the assumption that the service time vectors are i.i.d. Formula (30) for γ follows from the ergodic theorem for regenerative processes (see the formula in Cor.1 [4]). □

4.2.1.2 Tahoe (G1)-(G3)

The assumptions concerning the feedbacks are (G1) and (G3), with $RTO(n) = RTO$; we also assume that the service times can take a finite number of rational values. Under these assumptions, the variables $\{y_K(n) - y_K(n-w_{n-1}), n \in \mathbb{N}\}$ can only take a finite number of values too, say in a set Ψ, and this sequence has the same regenerative structure as above. The joint law of regeneration cycle T_1 and the dater $\overline{Z}(n)$ can be explicitly computed by the following recursion:

$$\begin{aligned}
&P(T_1 > n, \overline{Z}(n) = \overline{z})\\
&= P(\cap_{k=1}^{n}\{y_K(k) - y_K(k\text{-}w_{k\text{-}1}) < \mathrm{RTO}\}, \overline{Z}(n) = \overline{z})\\
&= \sum_{\overline{z}' \in \Psi} P(\cap_{k=1}^{n}\{y_K(k) - y_K(k\text{-}w_{k\text{-}1}) < \mathrm{RTO}\},\\
&\qquad \overline{Z}(n-1) = \overline{z}', \overline{Z}(n) = \overline{z})\\
&= \sum_{\overline{z}' \in \Psi} P(\overline{Z}(n) = \overline{z}, y_K(n) - y_K(w_{n\text{-}1}) < \mathrm{RTO} \mid \overline{Z}(n-1) = \overline{z}')\\
&\qquad P(T_1 > n-1, \overline{Z}(n-1) = \overline{z}').
\end{aligned}$$

This is valid for $n < b^*$. From this, one can derive a formula for the throughput using the ergodic theorem for regenerative processes in the same way as above:

$$\lambda = \frac{\sum_{k=0}^{b^*-1} P(T_1 > k)}{\sum_{k=0}^{b^*-1}\sum_{\overline{z} \in \Psi} P(T_1 > k, \overline{Z}(k) = \overline{z})\alpha(\overline{z})}, \tag{32}$$

where $\alpha(\overline{z})$ is the function defined in §4.1.

4.2.2 Extensions

A similar formula can be obtained for TCP Tahoe with slow start or for TCP Reno, and also for various extensions of the above model including independent packet losses as in §4.1.

Example 3. *Here we consider TCP Tahoe with $K = 3$ routers, under (G2) with $\beta_1 = \beta_3 = \infty$ and $\beta_2 = 3$, $w^* = 50$. The random variables $\sigma_i(n)$ are i.i.d. multinomial with the following values: $\sigma_1(n)$ is equal to $\{1,10,20\}$ with probability $0.1, 0.2, 0.7$; $\sigma_2(n)$ is equal to $\{13,15,17\}$ with probability $0.25, 0.5, 0.25$; $\sigma_3(n)$ is equal to $\{1,10,20\}$ with probability $0.7, 0.2, 0.1$.*

One of the curves of Figure 4 shows the evolution of $W(n)$ for TCP Tahoe model with slow start under these assumptions, whereas the other curve gives the evolution of $W(n)$

when exchanging the statistics of $\sigma_1(n)$ *and* $\sigma_3(n)$.

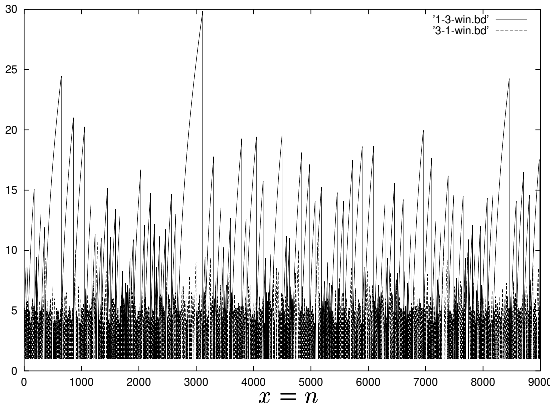


Figure 4: Evolution of $W(n)$

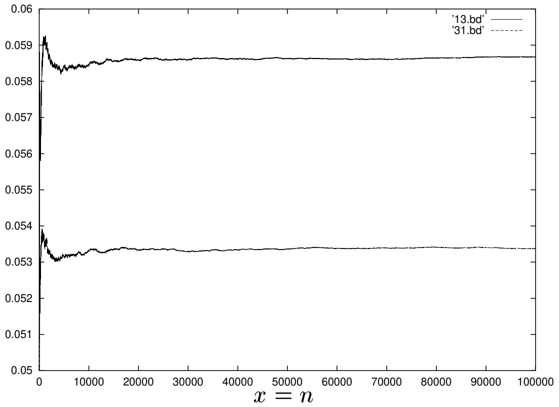


Figure 5: Comparison of Throughput

Figure 5 shows the comparison of the throughput of these two TCP Tahoe models. The first model gives a throughput of 0.058679 *(simulation of* 10^7 *packets), whereas that of the second one is* 0.053357*. So, the permutation of the characteristics of two routers may influence the value of the throughput: we cannot reduce the network to a single bottleneck router since the throughput may depend on the position of the bottleneck along the path.*

In the same way, mean values are not sufficient to predict the throughput: for instance, for the first model, when moving to deterministic service times equal to the mean values of the corresponding multinomial distributions, we find a throughput of 0.062112*, whereas when increasing variance of service times in routers 1 and 2 (*$\sigma_1(n)$ *equal to* $\{0, 32.2\}$ *with probability* $0.5, 0.5$ *and* $\sigma_2(n)$ *equal to* $\{0, 30\}$ *with probability* $0.5, 0.5$*), the throughput collapses to* 0.043819 *(*-30%*).*

5. FURTHER EXPLOITATION OF THE APPROACH

In the previous sections, we limited ourselves to the mean value of the saturated throughput. In fact, one can derive further results from our analysis, either analytically, or via our fast algebraic simulation algorithm; this concerns for instance:

1. The law of the instantaneous throughput $\lim_{n\to\infty} P((y_K(n+1)-y_K(n))^{-1} \leq x)$, which, under the setting of §4.1, is equal to $\sum_{(w,\overline{z})\in\Delta} \mathbb{1}_{\{\alpha(\overline{z})\geq 1/x\}}\pi(w,\overline{z})$; this is an important quantity which defines a natural indicator of QoS in complement to the average value;
2. The law of the end-to-end delay: $\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} P(y_K(n) - y_K(n - w_{n-1}) \leq x)$;
3. The law of the time D needed to transmit a file of size F; in first approximation, this law is given by the relation: $P(D > t) = P(y_K(F) > t)$, although more precise formulas can be derived when taking retransmissions of lost packets into account.

Within this framework, we could also possibly handle

4 Open model (i.e. non saturated source models like http sources), where the arrival process is described by its statistical characteristics; in this case, the basic equations read:

$$y_0(n) = [y_K(n - v_{n-1}) \otimes d_{K,0}] \oplus u(n),$$
$$y_i(n) = [y_{i-1}(n) \otimes d_{i-1,i} \oplus y_i(n-1)] \otimes \sigma_i(n),$$

$i = 1, \ldots, K$, where $u(n)$ denotes the time when packet n becomes available at the source node. This leads to a max-plus affine dynamical system where (5) has to be replaced by:

$$Z(n) = A_{v_{n-1}}(n) \otimes Z(n-1) \oplus V(n), \qquad (33)$$

where $V(n)$ is a vector built from $\{u(k)\}$ (see [3]).

5 Multiple connections cases allowing one to study interactions between several customers;

6 Multicast connections over a network involving a tree rather than a linear sequence of routers in series (see [5] for the constant window case).

7 Equation based control as considered in [7].

These last questions will be the object of future research.

6. CONCLUSION

We have shown that both in the saturated and the non saturated case, the adaptive feedback mechanism of TCP is a linear feedback in the max-plus algebra. This leads to a simple representation of the effect of this protocol on any network which admits a max-plus representation without the control, like tandem queues or the fork-join queue networks that one finds in multicast trees. We have deduced from this simple formulas for various deterministic service time models that refine well known results of the literature. These formulas confirm that in this case, the throughput only depends on the RTT and the bottleneck router rate, at least in the (G1) case. New formulas are also obtained for the random service time case, where the randomness is a natural way of representing the effect of the rest of traffic on the controlled connection. It is shown that in this case, one cannot obtain the throughput from mean values only, and that the order and the fine statistical behavior of the routers cannot be ignored. The set of all possible models within this setting is quite rich. One can indeed select a deterministic or random service time model, a congestion or loss based flow control; losses may stem from congestion, or timeouts, or be random, or any combination of the three; Reno or Tahoe can be selected, with or without slow start etc. We have shown how our approach could be used to analyze some of these combinations; we find it useful to stress that all such combinations can in principle be analyzed within this setting, which will be the object of our future research. More generally, this approach provides a generic framework for the simulation of TCP and related protocols over possibly large networks, based on simple algorithms with a low computational cost.

7. APPENDIX

7.1 Proof of Theorem 2

For all integers $1 \leq u, s \leq w^*$, and all vectors $z \in \mathbb{R}^{Kw^*}$ let $\{\tilde{W}(n), \tilde{Z}(n)\}$ be the sequence defined by

$$\begin{aligned} Z(n) &= A_{\lfloor W(n-1) \rfloor}(n) \otimes Z(n-1), \\ W(n) &= F\left(s, W(n-1), 1\right), \end{aligned}$$

with initial conditions $\tilde{W}(0) = u$, $Z(0) = z$ and $\Theta(0) = s$ (this is the sequence where we enforce $\Theta(n) \equiv s$ and $ACK(n) \equiv 1$). Let $\tilde{n}$ be the first integer n such that either loss or timeout are detected for packet n in the $\{\tilde{W}(n), \tilde{Z}(n)\}$ sequence, when making use of rule (6). In case no such event occurs, this means that the reference window size eventually stays constant and equal to W^*. If not, let $B^* = \tilde{W}(\tilde{n})$. Here are a few examples:

- Under (G1), $b^* = \lfloor B^* \rfloor$ is given by
$$b^* = \min\{n : n\sigma^* > \mathrm{S}\} = \left\lfloor \frac{\mathrm{S}}{\sigma^*} \right\rfloor + 1.$$
Note that in this case, $b^* \leq K + 1$ and only depends on σ^* and S.
- Under (G2),
$$\tilde{n} = \inf\{n : A_{\tilde{w}_{n-1}} \otimes \cdots \otimes A_{\tilde{w}_0} \otimes z \in \mathcal{S}_2\},$$
where $\mathcal{S}_2 = \{Z \in \mathbb{R}^{Kw^*}$ s.t. for some i, $1 \leq i \leq K$, $Z_{i-1} + d_{i-1,i} < Z_{i+\beta_i K}\}$. Note that in this case, b^* depends in general on the whole vectors $\sigma = (\sigma_1, \ldots, \sigma_K)$ and $\beta = (\beta_1, \ldots, \beta_K)$.
- Under (G3), the condition defining $\tilde{n}$ is as for (G2) but with, in place of $\mathcal{S}_2$, the set $\mathcal{S}_{3,n} = \{Z \in \mathbb{R}^{Kw^*}$ s.t. $Z_K - Z_{K+\tilde{w}_{n-1}K} > RTO(n)\}$.

Departing from $u = u_0 = 1$, $s = s_0 = \lfloor \alpha W^* \rfloor$, and $z = z_0 = (0, \ldots, 0)^t$, we either have a reference window size which eventually stays constant and equal to W^*; in this case, the result is proved with $a^* = b^* = w^*$. If not, after the value $B_0^* = B^*$ is reached, the window drops down, and it starts a new cycle similar to the first one, but this time with $u_1 = \lfloor \alpha B_0^* \rfloor$, $s_1 = s_0$ or $\lfloor \alpha B_0^* \rfloor$, and $z_1 = \tilde{Z}(\tilde{n})$. Here again, either $\tilde{W}(n) = W^*$ eventually, or we start a new cycle when the window size reaches the value B_1^* etc. In case the constant sequence with only W^* is never reached, there is an infinite sequence of such cycles that only differ in their initial conditions. Since the structure of the ith cycle is completely determined by the triple $(u_i, s_i, \overline{z}_i)$ where $\overline{z}_i$ is the class of z_i (see §2.6), and since there is only a finite set of possible values for such triples under our conditions, the periodicity follows. In case $RTO(n)$ is used in the initial condition, the proof can easily be adapted along the same lines. □

7.2 Proof of Corollary 1

We will only give the proof for the (G1) case. In view of Theorem 3, we have to check that the matrix $A_{b^*} \otimes A_{b^*-1}^{b^*-1} \otimes \cdots \otimes A_1$ has a unique eigenvalue and to compute this eigenvalue. The first property is immediate. The second one is proved via the following lemmas.

LEMMA 2. *For all $i \in \{1, .., b^* - 1\}$,*

$$A_i^i = \begin{array}{cl} \begin{pmatrix} M^i \oplus M' & MM' & \cdots & \cdots & M^{i-1}M' & \mathcal{E}\cdots \\ M^{i-1} & M' & \ddots & \ddots & M^{i-2}M' & \vdots \\ \vdots & \mathcal{E} & \ddots & \ddots & \vdots & \vdots \\ M^2 & \vdots & \ddots & M' & MM' & \vdots \\ M & \vdots & \vdots & \mathcal{E} & M' & \vdots \\ I_d & \mathcal{E} & \vdots & \vdots & \mathcal{E} & \vdots \\ \mathcal{E} & I_d & \mathcal{E} & \vdots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \end{pmatrix} & \begin{array}{l} \leftarrow ith \\ block \end{array} \\ \uparrow \; ith\ block & \end{array}$$

For all $i \geq 1$, $n \in \{1, .., i-1\}$,

$$A_i^n = \begin{array}{cl} \begin{pmatrix} M^n & \mathcal{E} \cdots \mathcal{E} & M' & \cdots & M^{n-1}M' & \mathcal{E}\cdots \\ \vdots & \vdots & \mathcal{E} & \ddots & \vdots & \vdots \\ M & \vdots & \vdots & \ddots & M' & \vdots \\ I_d & \mathcal{E} \cdots \mathcal{E} & \vdots & \vdots & \mathcal{E} & \vdots \\ \mathcal{E} & I_d\ \mathcal{E} \cdots \mathcal{E} & \vdots & \vdots & \vdots & \vdots \\ \vdots & \ddots\ \ddots\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} & \begin{array}{l} \leftarrow nth \\ block \end{array} \\ \uparrow \; ith\ block & \end{array}$$

PROOF. By finite induction. □

LEMMA 3. *For all $n \leq b^* - 1$, the matrix $M \otimes (M^n \oplus M') \otimes \cdots \otimes (M \oplus M')$ is irreducible and its eigenvalue is equal to $n\mathrm{S} + \sigma^*$.*

PROOF. $M \oplus M'$ is irreducible and this property is stable by left max-plus product by M, which implies the announced irreducibility property. Let $C^{(n)} = (M^n)_{1 \leq i \leq K, j=1}$, $C = C^{(1)}$. We have for all $n \geq 1$, for all $i \in \{1, .., K\}$,

$$\left(C^{(n)}\right)_i = (C)_i + (n-1) \max_{k=1,\ldots,i} \sigma_k.$$

and

$$M \otimes C^{(n)} = C^{(n+1)}, \; M' \otimes C^{(n)} = \mathrm{S} + (n-1)\sigma^* + C,$$

so that

$$(M^l \oplus M') \otimes C = C^{(l+1)} \oplus (\mathrm{S} \otimes C).$$

Since, for $n \in \{1,..,b^*-1\}$, $n\sigma^* \leq \mathrm{S}$, if $1 \leq l \leq b^*-1$, we have

$$(M^l \oplus M') \otimes C = \mathrm{S} + C,\ (M \oplus M') \otimes C^{(2)} = \mathrm{S} + \sigma^* + C.$$

Therefore $M \otimes (M^n \oplus M') \otimes \cdots \otimes (M \oplus M') \otimes C^{(2)} = n\mathrm{S} + \sigma^* + C^{(2)}$. Hence $C^{(2)}$ is an eigenvector of $M \otimes (M^n \oplus M') \otimes \cdots \otimes (M \oplus M')$ for the eigenvalue $n\mathrm{S} + \sigma^*$. $\square$

7.2.1 Proof of Corollary 1

We have

$$A_1 \otimes ((C^{(2)})^t, \cdots)^t \leq (\mathrm{S} + \sigma^* + (C)^t, \sigma^* + (C)^t, \cdots)^t$$

and

$$\begin{aligned} &A_2^2 \otimes (\mathrm{S} + \sigma^* + (C)^t, \sigma^* + (C)^t, \cdots)^t \\ &\leq (2\mathrm{S} + \sigma^* + (C)^t, 2\mathrm{S} + (C)^t, \mathrm{S} + \sigma^* + (C)^t, \cdots)^t, \end{aligned}$$

where $((C^{(2)})^t, \cdots)$ or $(\mathrm{S} + \sigma^* + (C)^t, \sigma^* + (C)^t, \cdots)$ are line vectors of dimension $K.w^*$ and where $\cdots$ are entries of these vectors that have no influence on the computation (for instance put $-\infty$).

Using Lemma 2 and the fact that for all $n < b^*$, $n\sigma^* \leq \mathrm{S}$, we get by induction that for all $i \in \{2,..,b^*-1\}$,

$$\begin{aligned} &A_i^i \otimes \big((i-1)\mathrm{S} + \sigma^* + (C)^t, (i-1)\mathrm{S} + (C)^t, \ldots, \\ &\quad (i-1)\mathrm{S} - (i-3)\sigma^* + (C)^t, (i-2)\mathrm{S} + \sigma^* + (C)^t, \cdots\big)^t \\ &\leq \big(i\mathrm{S} + \sigma^* + (C)^t, i\mathrm{S} + (C)^t, \ldots, \\ &\quad i\mathrm{S} - (i-2)\sigma^* + (C)^t, (i-1)\mathrm{S} + \sigma^* + (C)^t, \cdots\big)^t. \end{aligned}$$

Therefore

$$\begin{aligned} &\Big(A_{b^*} \otimes A_{b^*-1}^{b^*-1} \otimes \cdots \otimes A_1 \otimes ((C^{(2)})^t, \cdots)^t\Big)_{1 \leq i \leq K} \leq \\ &\quad ((b^*-1)\mathrm{S} + \sigma^*) \otimes (M \otimes C) \oplus \\ &\quad ((b^*-1)\mathrm{S} + (b^*-2)\sigma^*) \otimes (M' \otimes C) \\ &\leq (b^*-1)\mathrm{S} + \sigma^* + C^{(2)}. \end{aligned}$$

For $n < b^*$, let $B(n)$ denote the matrix $(A_{n+1} \otimes A_n^n \otimes \cdots \otimes A_1)_{1 \leq i,j \leq K}$. For all $n < b^*$, we have $B(n) \geq M \otimes (M^n \oplus M') \otimes \cdots \otimes (M \oplus M')$, so that

$$\begin{aligned} &\Big(A_{b^*} \otimes A_{b^*-1}^{b^*-1} \otimes \cdots \otimes A_1 \otimes ((C^{(2)})^t, \cdots)^t\Big)_{1 \leq i \leq K} \\ &\quad \geq M \otimes (M^n \oplus M') \otimes \cdots \otimes (M \oplus M') \otimes C^{(2)} \\ &\quad = (b^*-1)\mathrm{S} + \sigma^* + C^{(2)}, \end{aligned}$$

where the last equality follows from Lemma 3. Hence

$$\begin{aligned} &\Big(A_{b^*} \otimes A_{b^*-1}^{b^*-1} \otimes \cdots \otimes A_1 \otimes ((C^{(2)})^t, \cdots)^t\Big)_{1 \leq i \leq K} \\ &= (b^*-1)\mathrm{S} + \sigma^* + C^{(2)}. \end{aligned}$$

The relation $B(n) \geq M \otimes (M^n \oplus M') \otimes \cdots \otimes (M \oplus M')$ also implies that $B(b^*-1)$ is irreducible. Therefore the eigenvalue of $B(b^*-1)$, that is γ, is equal to $(b^*-1)\mathrm{S} + \sigma^*$.

8. REFERENCES

[1] E. Altman, J. Bolot, P. Nain, D. Elouadghiri, M. Erramdani, P. Brown, and D. Collange. Performance modeling of TCP/IP in a wide-area network. *INRIA Report*, March 1997.

[2] F. Baccelli and T. Bonald. Window flow control in FIFO networks with cross traffic. *Queueing Systems*, 32:195–231, 1999.

[3] F. Baccelli, G. Cohen, G. Olsder, and J. Quadrat. *Synchronization and Linearity.* Wiley, 1992.

[4] F. Baccelli, S. Gaubert, and D. Hong. Representation and expansion of (max,plus)–Lyapunov exponents. In *Proc. of 37th Annual Allerton Conf. on Communication, Control and Computing*, September 1999.

[5] A. Chaintreau, F. Baccelli, and C. Diot. Impact of network delay variation on multicast sessions performance with TCP-like congestion control. Preprint, 2000.

[6] C. Chang. *Performance guarantees in Communication Networks.* Springer Verlag, 1999.

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

[8] P. Hurley, J. Le Boudec, and P. Thiran. A note on the fairness of additive increase and multiplicative decrease. *Proceedings of ITC-16*, June 1999.

[9] S. Keshav. *An engineering approach to computer networking.* Addison–Wesley Professional Computing Series, 1997.

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

[11] M. Mathis, J. Semske, J. Mahdavi, and T. Ott. The macroscopic behavior of the TCP congestion avoidance algorithm. *Computer Communication Review*, 27(3), July 1997.

[12] J. Padhye, V. Firoiu, and D. Towsley. A stochastic model of TCP Reno congestion avoidance and control. *Technical Report CMPSCI, Univ. of Massachusetts*, 1999.

[13] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose. Modeling TCP throughput: a simple model and its empirical validation. *Proc. of ACM SIGCOMM*, 1998.

[14] R. Stevens. *TCP/IP Illustrated*, volume 1. Addison Wesley, 1994.

[15] G. Wright and R. Stevens. *TCP/IP Illustrated*, volume 2. Addison Wesley, 1995.