Interaction of TCP Flows as Billiards

François Baccelli, Dohy Hong

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Interaction of TCP Flows as Billiards

François Baccelli — Dohy Hong

N° 4437

April 2002

THÈME 1

ISSN 0249-6399 ISRN INRIA/RR--4437--FR+ENG

Interaction of TCP Flows as Billiards

François Baccelli [*] , Dohy Hong [†]

Thème 1 — Réseaux et systèmes
Projet TREC

Rapport de recherche n° 4437 — April 2002 — 38 pages

Abstract: The aim of this paper is to analyze the performance of a large number of long lived TCP controlled flows sharing many routers (or links), from the knowledge of the network parameters (capacity, buffer size, topology) and of the characteristics of each TCP flow (RTT, route etc.) in the presence of synchronization. This work is based on the AIMD model which describes the joint evolution of the window sizes of all flows in the congestion avoidance phase over a single bottleneck router, in terms of iterates of random affine maps. It is shown that the generalization of this dynamics to a network composed of several routers can be described in terms of iterate of random piecewise affine maps, or geometrically as a billiards in the Euclidean space with as many dimensions as the number of flow classes and as many reflection facets as there are routers. This can first be used as a simulation tool allowing one to emulate the interaction of millions of flows on tens of thousands of routers. This representation also leads to results of mathematical nature: this class of billiards exhibits both periodic and non-periodic asymptotic oscillations (to be interpreted as network level fluctuations for traffic aggregates), the characteristics of which are extremely sensitive to the parameters of the network; the consequences on TCP's fairness are exemplified on a few typical cases of small dimension. Finally, we also show that aggregated traffic generated by this billiards representation exhibits the same short time scale statistical properties as those observed on real traces.

Key-words: TCP, additive increase–multiplicative decrease algorithm, synchronization, fairness, IP traffic, dynamical system, billiards, iterates of piecewise affine maps, wavelet, fractal, product of random matrices.

[*] INRIA-ENS, ENS, 45 rue d'Ulm 75005 Paris, France {Francois.Baccelli@ens.fr}
[†] INRIA-ENS, ENS, 45 rue d'Ulm 75005 Paris, France {Dohy.Hong@ens.fr}

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 — Télécopie : +33 1 39 63 53 30

Interaction de Flots TCP et Billards

Résumé : Cet article analyse les performances d'un grand nombre de connexions de longue durée contrôlées par TCP et partageant un ensemble de routeurs et de liens, en présence de synchronisation entre les connexions. Les données du problème sont les paramètres du réseau (capacité des liens et des routeurs, taille des mémoires, topologie) et ceux de chaque connexion TCP (route, RTT etc.). Cette étude est fondée sur le modèle AIMD qui décrit l'évolution jointe des débits obtenus par des connexions, toutes dans la phase d'évitement de congestion et qui se partagent un seul routeur ou lien, au moyen d'itérées de fonctions aléatoires affines. Nous montrons comment généraliser ce type de dynamique à un réseau composé de plusieurs routeurs. Cette généralisation peut être décrite comme l'itération de fonctions aléatoires affines par morceaux, ou encore géométriquement comme un billard. L'espace d'état de ce billard est un polyèdre de l'espace euclidien, qui a pour dimension le nombre des flots et qui possède autant de faces qu'il y a de routeurs ou de liens dans le réseau. Cette représentation peut tout d'abord être utilisée comme un outil de simulation permettant d'émuler l'interaction de millions de connexions sur des réseaux composés de dizaines de milliers de routeurs. Elle conduit aussi à plusieurs résultats de nature mathématique: les oscillations asymptotiques de cette classe de billards (c'est à dire les fluctuations du trafic agrégé causées par le réseau) peuvent être périodiques ou non-périodiques et sont extrêmement sensibles aux paramètres du réseau; l'étude du régime stationnaire permet d'analyser le partage de la bande passante réalisé par TCP et notamment de quantifier l'équité du protocole. On en déduit aussi des modèles de trafic agrégé dont les propriétés statistiques de régularité locale sont similaires à celles observées sur le trafic réel.

Mots-clés : TCP, contrôle de congestion, contrôle de flux, algorithme des accroissements additifs et de la décroissance multiplicative, synchronisation, équité, trafic IP, système dynamique, billard, itération de fonctions linéaires par morceaux, ondelette, fractale, produit de matrices aléatoires.

1 Introduction

The AIMD model [4] describes the joint evolution of the congestion window size of N long lived (FTP or Peer to Peer type) flows controlled by TCP and sharing a single router or link, in terms of products of random matrices. The associated large population asymptotic model which concerns the case $N \to \infty$ was studied in [13].

The present paper studies the case when the TCP flows are heterogeneous (different RTT or routes) and when each flow goes through a route made of several tail-drop routers (throughout the paper, we will consider routers to be the possible bottlenecks; this could be replaced by links everywhere without altering the conclusions) in series. The corresponding model, which is introduced in §2, will be referred to as the multi-AIMD model.

Our aim to estimate the throughput obtained by each individual flow under the competition rules imposed by TCP, and also the fluctuations of this throughput, from the sole knowledge of the route and the RTT of each flow, and the characteristics of each router and link (buffer size, link capacity etc.) in the network.

This is of course related to the classical relationships that have been obtained between the packet loss probability and TCP throughput for a given session (see e.g. [20], [23]); in particular, it was shown in [4] that the single router AIMD dynamics resulted in a dependency between these quantities that was compatible with these formulas and was actually refining them in that it allowed one to assess the influence of synchronization.

The first models for the several router TCP network case are those of [15] and [10]. These papers analyzed the bandwidth sharing of different TCP flows over large networks in terms of optimization problems, and triggered a large number of further studies (see e.g. [19], [17], [18]). The prediction of the throughput in the several AQM router case has also been investigated via fixed point approximation methods for mean values in [7]. The approach that is proposed in the present paper addresses the same prediction question in the tail drop router case.

The main difference with these earlier approaches lies in the fact that we use a pathwise description of the dynamics of the interaction between flows, which takes into account discrete event phenomena that are of central importance for tail drop networks, such as congestion epochs, losses, synchronization of sources etc. and that allows one to analyze throughput fluctuations.

More precisely, the interaction is described by a set of evolution equations that generalize the random affine map description of the AIMD (one router) model. The basic multi-AIMD model can be seen as iterates of random *piecewise* affine maps. From this stochastic model, we define a large population asymptotic model. This asymptotic model can be seen as iterates of deterministic piecewise affine maps. These equations are shown to admit a geometrical representation in terms of a random or deterministic billiards in the Euclidean space. The dimension of this space is the number of different flow classes (typically, there is one flow class per route and RTT). This billiards has as many reflection facets as there are routers.

This new representation of the interaction between TCP flows and its exploitation are the main contributions of the present paper.

In §4, we show that this can be used as a simulation tool allowing one to emulate the interaction of millions of flows on tens of thousands of routers.

The billiards representation also leads to several results of mathematical nature which are gathered in §3. We first establish sufficient periodicity conditions for the asymptotic behavior of the throughput obtained by the interacting flows, as well as a conservation law that relates the intensities with which routers experience congestion. Billiards are known to possibly exhibit non-periodic asymptotic behaviors. We give evidence that this is possible for the class of billiards considered here. We also show

that this approach provides a new analytical and qualitative way of assessing TCP's fairness, which is also exemplified on a few cases of small dimension.

Finally, we study the statistical properties of aggregated traffic generated by this billiards representation and we show in §4.4 using wavelet tools that it exhibits the same short time scale statistical properties (fractal scaling) as those observed on real traces [24], [30], [9], [29], [2].

2 Notation and Model Description

2.1 Notation

The model parameters are the following:

- Network configuration: $\mathcal{R}$ is the set of routers; C_r is the capacity of router $r \in \mathcal{R}$; B_r is the buffer size of router $r \in \mathcal{R}$; all routers are assumed to be tail drop.
- Traffic configuration: $\mathcal{S}$ is the set of TCP flow classes; N_s is the number of TCP flows of class $s \in \mathcal{S}$; $\mathbb{P}_s$ is the route (forward and backward routes are assumed to be the same) of class s flows; depending on the circumstances, any such route will be considered as a sequence of routers or as a sequence of pairs of routers; $RTT_s = R_s$ is the propagation delay for class s flows, which is also the minimal RTT for this class; λ_s denotes the stationary throughput of class s (one of the key variables to be determined).
- Network and traffic configuration: $\mathcal{S}_r$ is the set of classes with a route using router r; M_r is the total number of flows sharing router r $M_r = \sum_{s \in \mathcal{S}_r} N_s$; for $s \in \mathcal{S}_r$, $a_{s,r} = N_s/M_r$, is the proportion of flows of class s within the set $\mathcal{S}_r$; $c_r = C_r/M_r$ is the throughput that one flow could get if the capacity were ideally and equally shared.

Assumption $\mathcal{A}$ (which will be used for in certain proofs) supposes that each router has at least one class with a route that contains this router only.

We now give the notation of the different state variables that we will use. Most of these variables refer to the sequence $\{T_n\}$ of all *congestion epochs* in the network. As in the AIMD model, T_n is the n-th epoch at which a loss (or several simultaneous losses) occur on at least one of the routers.

- $X^{(s,i)}(t)$ is the throughput of flow i of class s at time t;
- $X_n^{(s,i)} = X^{(s,i)}(T_n+)$ is the throughput of flow i of class s just after the n-th congestion time;
- $Y_n^{(s,i)} = X^{(s,i)}(T_n-)$ is the throughput of flow i of class s just before the n-th congestion time; by construction, it will always be true that for all $r \in \mathcal{R}$, and all time t,

$$\sum_{s \in \mathcal{S}_r} \sum_{i \in s} X^{(s,i)}(t) \leq C_r. \tag{1}$$

  The congestion epoch T_n will be said of type $r \in \mathcal{R}$, if $\sum_{s \in \mathcal{S}_r} \sum_{i \in s} Y_n(s,i) = C_r$. Nothing forbids to have T_n of both type r and r'.

- $\tau_{r,n+1}$ is the time between T_n and the next *virtual congestion epoch* of router r, which is defined as

$$\tau_{r,n+1} = \frac{C_r - \sum_{j,s \in \mathcal{S}_r} X_n^{(s,j)}}{\sum_{s \in \mathcal{S}_r} \frac{N_s}{R_s^2}};$$

this is the time that would elapse between T_n and the next congestion epoch on router r, should the capacities of all other routers be infinite;

- $\gamma_n^{(s,i,r)}$ is the multiplicative variable of flow $i \in s$ on router r : $\gamma_n^{(s,i,r)} = 1/2$ if there is a loss for flow i on router r; $\gamma_n^{(s,i,r)} = 1$ otherwise; so, $\gamma_n^{(s,i,r)} \equiv 1$ if $r \notin \mathbb{P}_s$.
  Throughout this paper, we will study several types of assumptions.
  The *rate-independent* (RI) model is that where the sequences $\gamma_n^{(s,i,r)}$ are
  - independent in r;
  - for all fixed r, independent and identically distributed (i.i.d.) in n;
  - for all fixed r and s, identically distributed and ergodic in $i \in s$;

  The *rate-dependent* (RD) case is that the law of $\gamma_n^{(s,i,r)}$ is a function of s, r and $Y_n^{(s,i,r)}$ (a flow of a given class that has a large instantaneous throughput has more chances to experience a loss than another flow of the same class but with a smaller throughput). Some RD cases will be introduced in §2.6 and studied in §3.3.
- $p_n^{(s,r)} = \mathbb{P}(\gamma_n^{(s,i,r)} = 1/2)$ is the *synchronization rate* of router r for the flows of class s at the n-th congestion epoch. In the rate-independent case, $p_n^{(s,r)} \equiv p^{(s,r)}$. The class-independent (CI) model is that where in addition, $p^{(s,r)} = p^{(r)}$, for all $s \in \mathcal{S}_r$.

The synchronization rate should not be confused with the packet loss rate. Let us illustrate this in the case of a single router under CI. Since the synchronization rate represents the *proportion* of flows that experience a loss during a congestion epoch, it is possible to simultaneously have a high synchronization rate and a low packet loss rate (e.g. when rarely all sources loose at the very same time) or the converse (e.g. when flows have frequent losses distributed like independent Poisson point processes). We show in §5.1 of the appendix how this synchronization rate can be estimated from the network parameters using simple queueing theoretic arguments. However, it should be noticed that the Multi-AIMD framework described above is not limited to the synchronization rate estimate proposed in §5.1, and that other and possibly better estimates could be used in place of it at later stages.

2.2 Dynamics in the Simplest Case

In a first step, it will be assumed that routers have no buffer capacity so that it makes sense to assume that the different RTTs are constant over time and equal to R_s for class s. We will see in §2.6 how to relax these assumptions that will help us simplifying the presentation of the dynamics.

Assume one knows $X_n^{(s,i)}$ for all i and s. Then the next inter-congestion time is $\tau_{n+1} = T_{n+1} - T_n = \min_{r\in\mathcal{R}} \tau_{r,n+1}$. Since, due to the AI rule, each flow of class s increases its send rate with slope $\frac{1}{R_s^2}$ (this is the slope obtained when assuming that the window size and the RTT are linked at any time by a Little like formula: $W = XR$), we get

$$Y_{n+1}^{(s,i)} = X_n^{(s,i)} + \frac{1}{R_s^2} \min_{r\in\mathcal{R}} \tau_{r,n} \tag{2}$$

$$= X_n^{(s,i)} + \frac{1}{R_s^2} \min_{r\in\mathcal{R}} \frac{C_r - \sum_{j,u\in\mathcal{S}_r} X_n^{(u,j)}}{\sum_{u\in\mathcal{S}_r} \frac{N_u}{R_u^2}}. \tag{3}$$

Let $r_n = \mathrm{argmin}_{r\in\mathcal{R}} \tau_{r,n}$. Assume that this set has one element. Then due to the MD rule,

$$X_{n+1}^{(s,i)} = \gamma_{n+1}^{(s,i,r_{n+1})} Y_{n+1}^{(s,i)}. \tag{4}$$

Should there be several elements in the last set, then one would apply the multiplicative rule for all routers of the set (the order in which the multiplicative decrease is made does not affect the result). So the global dynamical system reads: $\forall (i,s)$,

$$X_{n+1}^{(s,i)} = \gamma_{n+1}^{(s,i,r_{n+1})}\Big(X_n^{(s,i)} + \frac{1}{R_s^2}\min_{r\in\mathcal{R}} \frac{C_r - \sum\limits_{j,u\in\mathcal{S}_r} X_n^{(u,j)}}{\sum_{u\in\mathcal{S}_r}\frac{N_u}{R_u^2}}\Big) \tag{5}$$

$$= \gamma_{n+1}^{(s,i,r_{n+1})} = \Big(X_n^{(s,i)} + \frac{1}{R_s^2}\min_{r\in\mathcal{R}} \frac{c_r - \sum\limits_{u\in\mathcal{S}_r}\frac{a_{u,r}}{N_u}\sum\limits_{j\in u} X_n^{(u,j)}}{\sum_{u\in\mathcal{S}_r}\frac{a_{u,r}}{R_u^2}}\Big).$$

We see that the vector of throughputs at time T_{n+1} is obtained from that at time T_n via a random map which is piecewise affine.

2.3 Large Population Asymptotics

When the population grows large, this model admits a deterministic asymptotic model that generalizes that of the single router case as defined in [13]. All variables of interest then depend on a parameter N that grows large. We assume in particular that for all s, $N_s[N] = n_s N$ and that for all r, $C_r[N] = c_r M_r[N]$, so that the proportions

$$a_{s,r} = \frac{n_s}{\sum_{u\in\mathcal{S}_r} n_u}$$

are kept for all s and r.

Theorem 1 *Suppose the losses are rate-independent. Assume in addition that for all s, the initial conditions $X_0^{(s,i)}[N]$ are such that for all (deterministic) sequences of subsets $\sigma[N]$ of the set of flows of class s with a cardinal $|\sigma[N]|$ that tends to ∞, the empirical mean $\frac{1}{|\sigma[N]|}\sum_{i\in\sigma[N]} X_0^{(s,i)}[N]$ converges almost surely (a.s) to a deterministic limit $x_0^{(s)}$ which does not depend on the sequence of subsets that is chosen. Then for all n,*

$$\exists \lim_{N\to\infty} \frac{1}{|\sigma[N]|}\sum_{i\in\sigma[N]} X_n^{(s,i)}[N] = x_n^{(s)} \quad a.s.$$

with $x_n^{(s)}$ deterministic, and such that the limit does not depend on the sequence of subsets that is chosen. In addition, the variables $x_n^{(s)}$, $s\in\mathcal{S}$, satisfy the evolution equation

$$x_{n+1}^{(s)} = \overline{\gamma}_{n+1}^{(s,\overline{r}_{n+1})}\left[x_n^{(s)} + \frac{1}{R_s^2}\overline{\tau}_{n+1}\right], \tag{6}$$

where $\overline{\gamma}_n^{s,r} = \mathbb{E}[\gamma_n^{(s,i,r)}]$,

$$\begin{aligned}\overline{\tau}_{n+1} &= \min_{r\in\mathcal{R}} \frac{c_r - \sum_{u\in\mathcal{S}_r} a_{u,r}x_n^{(u)}}{\sum_{u\in\mathcal{S}_r}\frac{a_{u,r}}{R_u^2}}\\ \overline{r}_{n+1} &= \operatorname{argmin}_{r\in\mathcal{R}} \frac{c_r - \sum_{u\in\mathcal{S}_r} a_{u,r}x_n^{(u)}}{\sum_{u\in\mathcal{S}_r}\frac{a_{u,r}}{R_u^2}}.\end{aligned}$$

If in addition, the initial condition is such that for all (s,i), the a.s. limit $\lim_{N\to\infty} X_0^{(s,i)}[N] = X_0^{(s,i)}[\infty]$ exists, then for all n, the a.s. limit $\lim_{N\to\infty} X_n^{(s,i)}[N] = X_n^{(s,i)}[\infty]$ also exists, and the sequence of random variables $X_n^{(s,i)}[\infty]$ satisfies the following stochastic recurrence equation:

$$X_{n+1}^{(s,i)}[\infty] = \gamma_{n+1}^{(s,i,\overline{r}_{n+1})}\left[X_n^{(s,i)}[\infty] + \frac{1}{R_s^2}\overline{\tau}_{n+1}\right], \tag{7}$$

with $\overline{r}_{n+1}$ and $\overline{\tau}_{n+1}$ the variables defined in the last deterministic equations.

The proof is forwarded to the appendix (§5.2). In this last model, we will denote by $y_n^{(s)}$ (resp. $x^{(s)}(t)$) the variables defined as $x_n^{(s)}$ but from the random variables $Y_n^{(s,i)}$ (resp. $X^{(s,i)}(t)$). We deduce the following inequalities from (1): for all t and r, $\sum_{s\in\mathcal{S}_r} a_{s,r}x^{(s)}(t) \le c_r$.

In what follows, (5), satisfied by the actual throughput vector, will be referred to as the *stochastic multi-AIMD model* and (6), satisfied by the the vector of empirical means, will be referred to as the associated *large population asymptotic model*.

An important question (that will not be discussed here) is that of the speed of convergence and of the nature of the error term when approximating the model with N large but finite by the asymptotic model. First results on the convergence of the moments are reported in [13] for the single router case. In many mean field models, a central limit theorem can be established, which allows one to prove that the fluctuations around the limit are Gaussian, and to estimate them (see e.g. [11]). Whether this type of results also holds for the general class of dynamics identified here will be the object of future research.

Equivalent large population equations

If we take as state variables $\widetilde{x}_s = n_s x_s$, in place of x_s, $s \in \mathcal{S}$, then the large population equations can be rewritten under the equivalent form, which will also be used later:

$$\widetilde{x}_{n+1}^{(s)} = \overline{\gamma}_{n+1}^{(s,\widetilde{r}_{n+1})}\left[\widetilde{x}_n^{(s)} + \frac{1}{\widetilde{R}_s^2}\widetilde{\tau}_{n+1}\right], \tag{8}$$

where $\widetilde{R}_s = R_s/\sqrt{n_s}$ and where

$$\begin{aligned}
\widetilde{\tau}_{n+1} &= \min_{r\in\mathcal{R}} \frac{(\sum_{s\in\mathcal{S}_r} n_s)c_r - \sum_{u\in\mathcal{S}_r}\widetilde{x}_n^{(u)}}{\sum_{u\in\mathcal{S}_r}\frac{1}{\widetilde{R}_u^2}} = \overline{\tau}_{n+1} \\
\widetilde{r}_{n+1} &= \operatorname{argmin}_{r\in\mathcal{R}} \frac{(\sum_{s\in\mathcal{S}_r} n_s)c_r - \sum_{u\in\mathcal{S}_r}\widetilde{x}_n^{(u)}}{\sum_{u\in\mathcal{S}_r}\frac{1}{\widetilde{R}_u^2}} = \overline{r}_{n+1}.
\end{aligned}$$

2.4 The Three Levels

The proposed approach allows one to decouple three different levels:

1. The *network level*, which is captured by (6) or (8), and where the large population averaging takes place; this level, which we believe to be the main new paradigm identified in the present paper, will be the central object of the mathematical study of the present section.

2. The *flow level*, which is captured by the stochastic equation (7); the results obtained at the network level (e.g. the determination of the period of the sequence $\{\overline{\tau}_n, \overline{r}_n\}$, see Theorem 1) can be used to determine the effect of the network on each flow via the stochastic recurrence (7); this type of stochastic recurrences was already studied (at least in some special cases) in [4], [13] and more recently in [8], and we will not pursue the mathematical analysis of this level in the present paper.

3. The packet level, which is captured by the synchronization rate formula (19); the delay of reaction proper to TCP is present and taken into account via the synchronization rate formula (see Theorem 2). This packet level influences both the network and the flow levels via the impact of the synchronization rate on this level.

As we will see, each level is responsible for parts of the fluctuations of the throughput obtained by flows of aggregates of flows. By decoupling, we simply mean that the fluctuations of all levels can be analyzed independently.

2.5 Billiards Interpretation

The dynamics of the large population asymptotic model can be seen as that of a deterministic billiards model, the geometry of which is determined by the routes and the RTTs of the various flow classes and the capacity of the routers. The stochastic multi-AIMD model can be seen as a randomized version of the billiards.

This is illustrated by the three–class, two–router network of Figure 1. Here $c_1 = c_2 = \frac{C}{2}$,

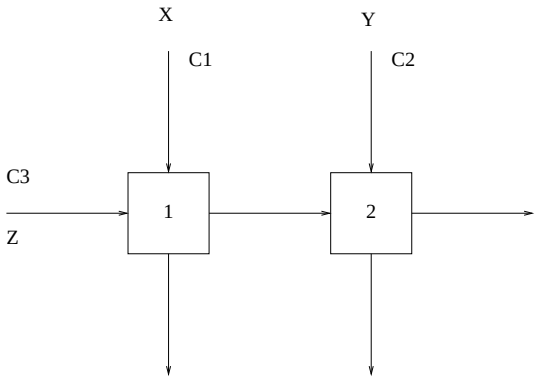

Figure 1: 2 Router, 3 Class Network Topology

$\mathcal{S}_1 = \{1, 3\}$, $\mathcal{S}_2 = \{2, 3\}$, and $a_{s,r} = \frac{1}{2}$ (or equivalently $n_s = 1$) for all s and r. As for RTT's, we take $R_s = 1$ for all s. The synchronization rates are all assumed to be equal to 1, so that $\overline{\gamma}_n^{s,r} = 1/2$ for all s and r. Let us look at the evolution of the large population asymptotic vector $(x^{(1)}(t), x^{(2)}(t), x^{(3)}(t))$ in the three dimensional Euclidean space (X, Y, Z). Notice that in this particular case (6) and (8) are the same. This vector lives in the polyhedron:

$$X \geq 0, \quad Y \geq 0, \quad Z \geq 0, \quad X + Z \leq C, \quad Y + Z \leq C,$$

which is depicted on Figure 2 and is the domain of the billiards. The plane H_1 ($X+Z = C$) represents the capacity constraint of router 1, with a similar interpretation for the plane H_2 ($Y + Z = C$). From

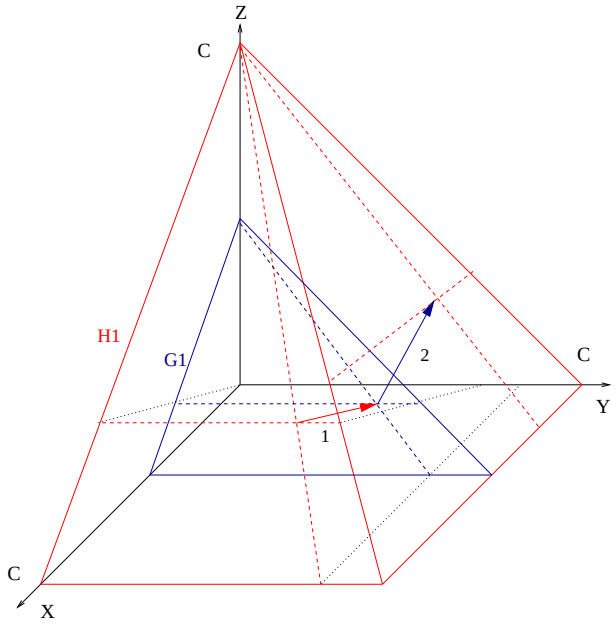

Figure 2: Billiards Domain

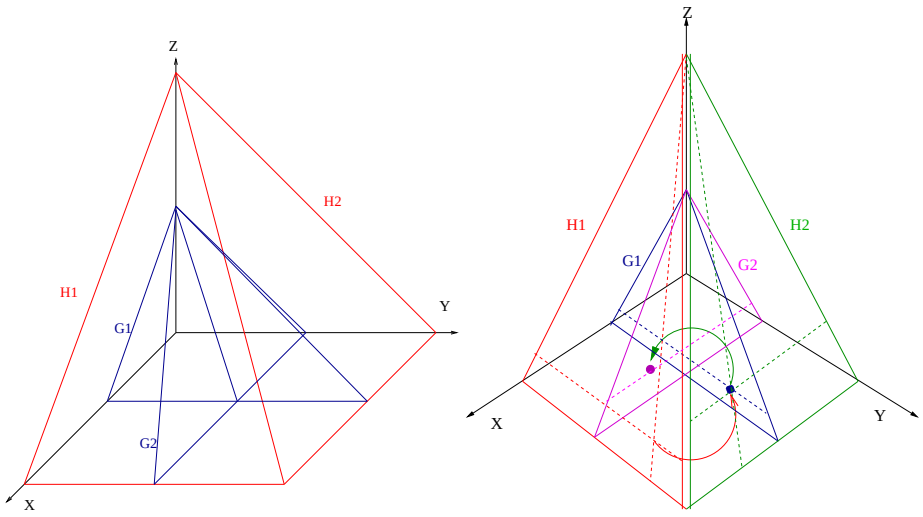

Figure 3: Billiards facets (left) and trajectories (right)

any point in the domain, the ball (i.e. the throughput process) moves linearly with time along the main diagonal with a constant velocity, as a consequence of the AI rule and the fact that all RTT's are the same. If the ball reaches the plane H_1, then it instantaneously jumps (red arrow, or arrow 1 for black and white reading) to the plane G_1 ($X + Z = C/2$), which describes the occurrence of losses on router 1. After this jump, the process (X, Y, Z) grows along the main diagonal again (blue arrow or arrow 2) until it hits one of the planes H_1 or H_2 (the last one is met first for this trajectory) and so on. The left part of Figure 3 gives a more complete view of the parts of all planes H_1, H_2 and G_1, G_2, where similar phenomena take place, namely jumps from H_2 to G_2 and growth along the

main diagonal from G_2 to either H_1 or H_2. The right part of Figure 3 depicts a (projective) view along the main diagonal. In this projective view, any linear increase is just a point. An instance of sequence of additive increases and multiplicative decreases (which appear as arrows) is illustrated there where the ball departs from H_1 and then successively hits G_1, H_2, G_2 and H_1.

In the stochastic model, the multiplicative dynamics is a randomized version of the last one: facets H_1 and H_2 still exist, but reflection on say H_1 projects the ball in a random neighborhood of the point of G_1 where the deterministic billiards jumps. The neighborhood in question is approximately Gaussian in the three dimensional space. In what follows, we will concentrate on the deterministic billiards model for mathematical analysis, and use the stochastic version for simulation studies. For relations between the two, see the comments at the end of §2.2 and 3.

2.6 Discussion and Model Refinements

The basic model has a certain number of weaknesses that are discussed below. We also show how to correct them. These corrections are easily taken into account in the simulation section, and as we will see, some of them are also analytically tractable.

2.6.1 Small Throughputs

It should first be noticed that a given window cannot be halved an unbounded number of times, be it only because it always remains larger than or equal to 1, whereas, in our model, one specific coordinate of X can actually be halved an unbounded number of times. We argue that although windows cannot be halved infinitely often indeed, throughputs (which are the actual state variables of the model) can become "multiplicatively small" in case of repeated timeouts. In such a case the doubling of the RTO variable (see e.g. [12]) has the very same qualitative effect on session inter-packet times and therefore throughput as the one of the aimd model. For more comments on the matter see [4].

2.6.2 Model with Finite Buffer Capacity

In the basic model, when for some r, $\sum_{s\in\mathcal{S}_r}\sum_{i\in s} X^{(s,i)}(t)$ reaches C_r, losses occur on this router. This is of course unrealistic as the router buffer size should be taken into account. Let B_r be the buffer size of router r. The equations can be adapted to take buffering into account. The simplest (heuristic) adaptation consists in replacing C_r by a random sequence $C_n^r = C_r + \kappa_n$, with $\kappa_n < \sqrt{2B_r\sum_{s\in\mathcal{S}_r}\frac{N_s}{R_s^2}}$. The justification is the following: if at time 0 the total arrival rate is C_r and the buffered fluid amount is 0, then at time t, in the absence of reaction from the flows, the total arrival rate is $C_r + t(\sum_{s\in\mathcal{S}_r} N_s/R_s^2)$ and the buffer contents is $t^2\sum_{s\in\mathcal{S}_r} N_s/2R_s^2$. Thus, when the buffer size B_r is reached, the total arrival rate is equal to $C_r + \sqrt{2B_r\sum_{s\in\mathcal{S}_r}\frac{N_s}{R_s^2}}$. However, if at the time the total arrival rate reaches C_r, the buffered fluid amount is larger than 0, then the equivalent capacity now depends on this amount and we have to take some random sequence κ_n which satisfies the last inequality. For more on the matter, see [8].

This way of representing the effect of buffers does not take into account the fact that RTTs increase with buffer contents. So, it can only be used whenever buffers are small enough for allowing one to neglect this increase of RTT. A better (though non-linear) way is proposed in the next subsection.

2.6.3 The Non-Linear AIMD Model

In the FIFO case, the following evolution equations should be used in-between congestion epochs:

$$\frac{dX^{(s,i)}(t)}{dt} = \frac{1}{(R_s(t))^2}, \quad R_s(t) = R_s + \sum_{r \in \mathbf{P}_s} \frac{B_r(t)}{C_r},$$
$$\frac{dB_r(t)}{dt} = \left(\sum_{s \in \mathbf{P}_r} \sum_{i=1}^{N_s} X^{(s,i)}(t) - C_r \right) 1_{B_r(t)>0},$$

with R_s the propagation delay (minimal value of RTT) for class s. The evolution equations at congestion epochs are exactly as in the basic Multi-AIMD model. Notice that the slow start phase can easily be represented by a slight adaptation on the non-linear dynamics.

2.6.4 Model with Rate-Dependent Losses

For a rate-dependent synchronization stochastic model, the law of the random variable $\gamma_n^{(s,i,r)}$ depends on the throughput just before the n-th congestion time that is on $Y_n^{(r,s,j)}$, $j \in s$, $r \in \mathbb{P}_s$. In the large population asymptotic model, $\overline{\gamma}^{(s,r)}$ depends on $y_n^{(r,s)}$, $r \in \mathbb{P}_s$. An example of this situation was studied in [13] in the single router case. Another somewhat simpler form is studied in §3.3 below, where the synchronization rate is supposed to be a function of the stationary throughput. Since the last quantity is unknown (this is actually the main quantity to be determined), this leads to fixed point problems. It should be possible to extend Theorem 1 to certain RD cases (as it in the single router case [13]).

3 Analysis of the Network Level Equations

Billiards models have been extensively studied using ergodic theory (e.g. Sinai's billiards [26]). Billiards reducible to iterates of piecewise affine maps (our TCP billiards belong to this class) have also been studied (see e.g. [22]). Within this piecewise affine class, even in the case where all maps are contracting, there are unfortunately no general results holding for all dimensions. In particular, there are simple examples of small dimension where some situations lead to a periodic behavior, whereas others lead to a non-periodic one. The subclass of TCP billiards (piecewise affine billiards that stem from TCP dynamics) has some specific properties that could make it amenable to a more specific analysis: the domains where the map is affine are intersections of half-spaces; each map is the composition of the multiplication by a diagonal matrix and of a projection on along some direction etc.

3.1 Periodic Regime

The aim of this subsection is to give a general sufficient condition for having only periodic behaviors; this sufficient condition is in term of a sequence of linear programs; as we will see, this also allows one to determine the period and the orbit.

In this subsection, we will assume Assumptions $\mathcal{A}$. This assumption is not essential for most properties. However, it ensures that each router reaches congestion infinitely often, which simplifies the exposition of the results.

The reference space is the Euclidean space of dimension K, where K is the cardinality of $\mathcal{S}$. We will use (8) rather than (6). We will drop the "tilde" on the variables for the sake of easy notation. Let

H_r denote the hyperplane

$$\sum_{s\in\mathcal{S}_r} x^{(s)} = \left(\sum_{s\in\mathcal{S}_r} n_s\right) c_r, \tag{9}$$

which will be referred to as facet r of the billiards.

3.1.1 Discrete Time Dynamics

Rather than the continuous time dynamics, we will study the *discrete time dynamics*, $\{y_n\}$, which gives the throughput process sampled *just before* congestion epochs, that is when the ball hits one of the facets.

For all r, let $\phi_r : \mathbb{R}^K \to \mathbb{R}$ denote the affine form

$$\phi_r(y) = \frac{c_r - \sum_{u\in\mathcal{S}_r} a_{u,r} y^{(u)}}{\sum_{u\in\mathcal{S}_r} \frac{a_{u,r}}{R_r^2}}.$$

For all r, s, let $F_{r,s}$ denote the subset of H_r where when applying the discrete time dynamics once, the ball hits facet s at next step. Since the open domain of $\mathbb{R}^K$ where s is hit before any other facet is that where $\phi_s(y) < \phi_v(y)$ for all $v \neq s$, each $F_{r,s}$ is a convex polyhedron of H_r which is the intersection of H_r and of a finite family of half spaces. By definition,

- on $F_{r,s}$, the one-step discrete time dynamics is some affine map that will be denoted by $B_{r,s}$;
- the family $F_{r,s}$, $s = 1, \ldots, |\mathcal{R}|$ =card($\mathcal{R}$) is a partition of H_r.

More generally, for all sequences $r_1, \ldots, r_k$ with elements in $\{1, \ldots, |\mathcal{R}|\}$, let $F_{r,r_1,\ldots,r_k}$ be the subset of H_r where when applying the discrete time dynamics k times, one successively visits the facets $r_1, \ldots, r_k$. For all sequences $r_1, \ldots, r_k$,

- on $F_{r,r_1,\ldots,r_k}$, the k-step discrete time dynamics is the affine map $B_{r,r_1,\ldots,r_k} = B_{r_{k-1},r_k} \circ \cdots \circ B_{r,r_1}$;
- the domain $F_{r,r_1,\ldots,r_k}$ is the (possibly empty) intersection of H_r and of the finite family of half spaces:

$$\begin{aligned}
\phi_{r_1}(y) &< \phi_v(y), \quad \forall v \neq r_1;\\
\phi_{r_2} \circ B_{r,r_1}(y) &< \phi_v \circ B_{r,r_1}(y), \quad \forall v \neq r_2;\\
\cdots &\quad \cdots\\
\phi_{r_k} \circ B_{r,r_1,\ldots,r_{k-1}}(y) &< \phi_v \circ B_{r,r_1,\ldots,r_{k-1}}(y), \quad \forall v \neq r_k;
\end{aligned}$$

- the family $F_{r,r_1,\ldots,r_k}$ $r_i = 1, \ldots, |\mathcal{R}|$, $i = 1, \ldots, k$ forms a partition of H_r.

The discrete time dynamics features a sequence $\{r_i\}$, $i \geq 0$ of faces that are successively hit, which depends on the initial condition for the throughput vector. Using the $\mathcal{A}$ assumption, one proves:

Lemma 1 *For all r, for all initial conditions in H_r, the sequence $\{r_i\}$, with $r_0 = r$, exits r in a number of steps bounded by a constant e_r. For all r, for all initial conditions in $F_r = \cup_{s\neq r} F_{r,s}$, the sequence $\{r_i\}$, with $r_0 = r$, returns to r in a number of steps bounded by a constant f_r.*

We will say that step i is an *exit step from* r if $r_i = r$ and $r_{i+1} = s \neq r$. Fix $r_0 = r$, some initial condition in F_r (so that $i = 0$ is an exit step from r) and consider the discrete time dynamics until the next exit step from r. This next exit step is finite as a corollary of Lemma 1.

The set of possible facet sequences $r_0 = r, r_1, \ldots, r_k$, $k \in \mathbb{N}$, r_1 in $\{1, \ldots, |\mathcal{R}|\}$, between two exit steps from r is that with

- $k \leq e_r + f_r$;
- $r_{k-1} = r$ and $r_l \neq r$ for all $l = 1, \ldots k$ with $l \neq k$.

The cardinality of this set is finite, and it will be denoted by q_r.

Denote by θ_r the mapping that associates to each initial condition in F_r the point where the ball is located at the next exit step from r, or equivalently when it first returns to F_r. We summarize what precedes in the following lemma:

Lemma 2 *The domain F_r can be partitioned into a finite number of convex polyhedrons $E_{r,1}, E_{r,2}, \ldots, E_{r,q_r}$, each of which is the intersection of H_r and of a finite family of half spaces. These domains, which we will refer to as the* linearity domains *of facet r, are such that for all n, for all initial conditions in the interior of $E_{r,n}$,*

- *the sequence of facets that are successively hit until the first return to F_r is exactly the same;*
- *the mapping θ_r is some affine mapping $A_{r,n}$ from F_r to itself.*

3.1.2 Sufficient Conditions for Facet Periodicity

Here is the simplest sufficient condition for the periodicity of the facet sequence, which will be referred to as the *inclusion test* in what follows:

Lemma 3 *If for some r, for all $n = 1, \ldots, q_r$, the set $A_{r,n}(E_{r,n})$ is completely included in one of the linearity domains, say $E_{r,g(n)}$, of facet r, then, for all initial conditions of the throughput process, the sequence of facets that are successively hit is ultimately periodic.*

Let Δ_r denote the finite set of all hyperplanes used in the definition of the linearity domains of facet r, namely:

$$\begin{aligned}
\phi_{r_1}(y) &= \phi_v(y), \quad \forall v \neq r_1; \\
\phi_{r_2} \circ B_{r,r_1}(y) &= \phi_v \circ B_{r,r_1}(y), \quad \forall v \neq r_2; \\
\cdots & \quad \cdots \\
\phi_{r_k} \circ B_{r,r_1,\ldots,r_{k-1}}(y) &= \phi_v \circ B_{r,r_1,\ldots,r_{k-1}}(y), \quad \forall v \neq r_k,
\end{aligned}$$

where $r_0 = r, r_1, \ldots, r_k$, ranges over the set of all possible facet sequences between two exit steps from r.

The condition of the last lemma can be rewritten as the solution of the linear problem

$$\begin{aligned}
x &\in A_{r,n}(E_{r,n}) \\
y &\in \delta_h,
\end{aligned}$$

being the empty set for all $n = 1, \ldots, q_r$ and all hyperplanes δ_h in Δ_r.

In case this condition is not satisfied, one ought to check whether $\theta_r^2 = \theta_r \circ \theta_r$ satisfies the appropriate inclusion property: denote by $E_{r,1}^{(2)}, \ldots, E_{r,q_r^{(2)}}^{(2)}$ the linearity domains of this map (again defined as the intersection of H_r and of certain half spaces) and by $A_{r,1}^{(2)}, \ldots, A_{r,q_r^{(2)}}^{(2)}$ the affine maps on these domains. Then if $A_{r,n}^{(2)}(E_{r,n}^{(2)}) \subset E_{r,g(n)}^{(2)}$ for all n, for some function $g : \{1, \ldots, |\mathcal{R}|\} \to \{1, \ldots, |\mathcal{R}|\}$, then for all initial conditions of the throughput process, the sequence of facets that are successively hit is ultimately periodic.

A similar sufficient condition (which will be referred to as the inclusion test of order k) can be obtained from $\theta_r^k = \theta_r \circ \theta_r^{k-1}$ for any $k \geq 2$.

3.1.3 Sufficient Conditions for Billiards Periodicity

The proof of the next lemma is forwarded to Appendix 5.3.

Lemma 4 *Let $r_1, r_2, \ldots, r_n$ be a fixed periodic sequence of facets. Then in the class independent (CI) case, for all dynamics with a sequence of facets which is ultimately periodic, with period $r_1, r_2, \ldots, r_n$, the associated billiards is asymptotically periodic and with a uniquely defined period that is independent of the chosen initial conditions.*

Combining this lemma and the sufficient conditions for the periodicity of the facet process provides a sufficient condition for the periodicity of the throughput process.

An interesting question which is still open at this stage is that of the irreducibility. When $\mathcal{A}$ is not assumed to hold, it is quite easy to find networks (e.g. with 3 routers) where two or more different periodic regimes can be reached depending on the initial condition. When $\mathcal{A}$ holds, we did not find situations with multiple non-degenerate periodic regimes yet (i.e. regimes where the periodic regime is such that ball bounces on the intersection of more than one facet during the period).

Notice that in the general case, the number of linearity domains grows in a non-polynomial way with K and the order k of the inclusion test. This clearly indicates that this method, when employed as an analytical modeling tool, can unfortunately not be used to assess the properties of large networks. As we will see below, it is however an efficient tool for analyzing small networks.

Example 1 Consider the network introduced at the end of the last section. The ball lives in the polyhedron of Figure 3. Let $y_n = (X_n, Y_n, Z_n)$ be the three dimensional vector of throughputs just before the n-th congestion epoch. At these epochs, the ball is on one of the two facets H_1 and H_2 (respectively the red or leftmost and the green or rightmost sides of this polyhedron). Let Δ be the dashed line on the red (leftmost) facet. This line partitions H_1 into two triangular domains, the rightmost of which is $F_{1,2}$: if y_0 belongs to $F_{1,2}$, y_1 belongs to H_2, and it is obtained from y_0 by the affine transformation

$$B_{1,2}(X,Y,Z) = \frac{1}{4}\begin{pmatrix} 2 & -2 & -1 \\ 0 & 2 & -1 \\ 0 & -2 & 1 \end{pmatrix}\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \frac{1}{2}\begin{pmatrix} C \\ C \\ C \end{pmatrix}.$$

If y_0 belongs to the complement of $F_{1,2}$, then y_1 belongs to H_1 and it is given from y_0 via another affine transformation. The situation is similar on the facet H_2, to which one associates two domains $F_{2,1}$ (which leads to H_1 via $B_{2,1}$) and its complement (which leads to H_2).

In this example F_1 has a single linearity domain $E_{1,1} = F_{1,2,1,2}$ and the inclusion test holds as $F_{1,2,1,2} \subset F_{1,2}$.

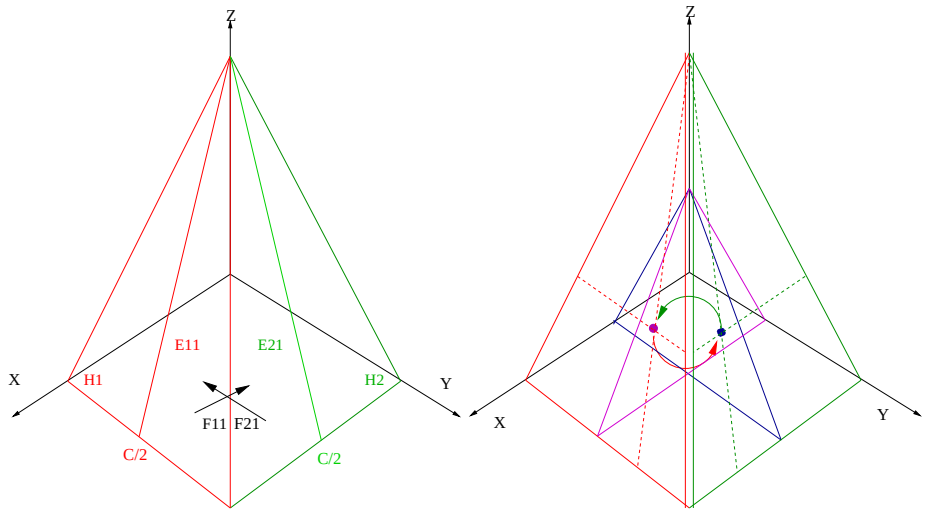

Figure 4: Periodic billiards trajectories

This implies that regardless of the initial condition in H_1, the sequence of facets is ultimately periodic with period 2. and that the sequence of affine operators that are applied is periodic too, also with period 2 Since $A_{1,1}$ is a contraction from $F_{1,1}$ to itself, it admits a unique fixed point.

It is not difficult to check that the last conclusions actually hold for any configuration with everything as above but for general RTTs R_x, R_y and R_z and general synchronization rates p_x, p_y and p_z provided that $R_x = R_y$ and $p_x = p_y$.

In the special case $R_x = R_y = R_z = 1$, $p_x = p_y = p_z = 1$ that is considered above, the fixed point of $A_{1,1}$ is easily computed as being $X^* = C/2, Y^* = 2C/3, Z^* = C/3$. This fully determines the periodic behavior which is unique in this case, at least when excluding degenerate periodic regimes such as the one that oscillates from a point of the line $X = Y = 1 - Z$ to another point of the line $X = Y = 1/2 - Z$.

The stationary throughput in continuous time, which is the average of the periodic throughput process depicted on the right part of Figure 4, is $\lambda_x = \lambda_y = C/2$ and $\lambda_z = C/4$.

In the case $R_x = R_y = 1$ and $p_x = p_y = p_z = 1$, direct calculations give

$$\lambda_x = \lambda_y = \frac{3CR_z^2}{2(2R_z^2+1)}, \quad \lambda_z = \frac{3C}{4(2R_z^2+1)}. \tag{10}$$

Example 2 We come back to the network of Figure 2, still with $n_s = 1$ for all s. Here, we take $c_1 = 2c_2 = 1/2$. The billiards associated with (8) now lives in a less symmetrical polyhedron depicted in Figure 5.

The linearity domains of H_1 are given in Figure 6, which gives a view of H_1 projected on the $X = 0$ plane. From $E_{1,1} = F_{1,2,2,1,2}$, the ball hits H_2 twice before coming back to F_1, whereas in $E_{1,2} = F_{1,2,2,2,1,2}$ it hits H_2 three times before returning to F_1. The Δ line that separates the two linearity domains is here $10Y + 11Z = 8$, $Z + X = 2$. In this case, the inclusion test does not hold for $k = 1$ but it does for $k = 6$, and there is a unique periodic regime of period 19.

The projection of this periodic regime on the $Y = 0$ plane is depicted on Figure 7. The leftmost region (up to 1.03 of the horizontal axis) is the image of $E_{1,2}$ by θ, whereas the region between 1.03

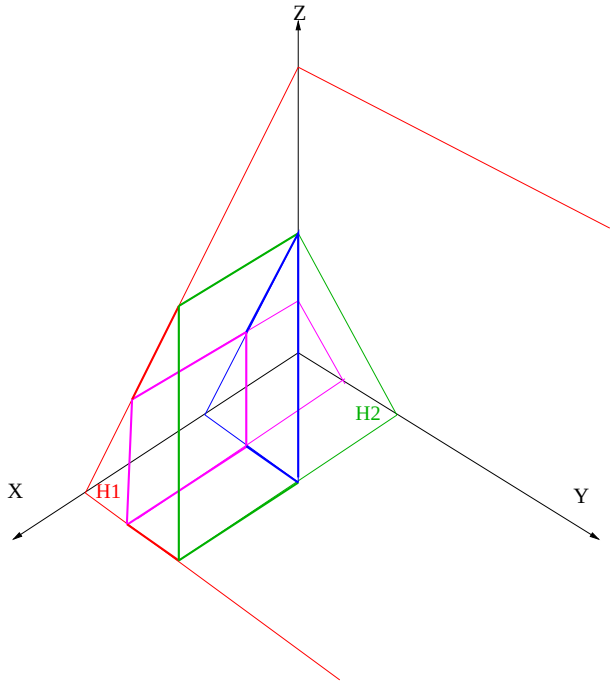

Figure 5: Non Symmetrical Case

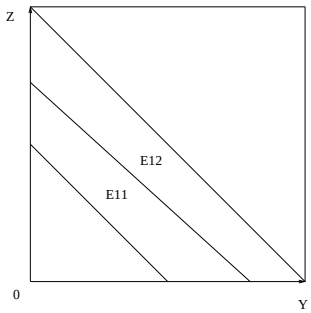

Figure 6: Linearity Domains

and 1.33 is the image of $E_{1,1}$. The rightmost line is the projection of H_1 on H_2. The facet period is

$$((H_2)^3 E_{1,2})^4 (H_2)^2 E_{1,1}$$

(with a notation that should be clear) and the billiards period is easily deduced from the unique fixed point of the corresponding affine operator.

3.2 Non-Periodic Regimes

The aim of this section is to provide numerical evidence that non-periodic facet sequences are possible. The way for searching for such behaviors consists in choosing some topology where a single parameter like e.g. the speed of some router is varied and in plotting the period of the billiards.

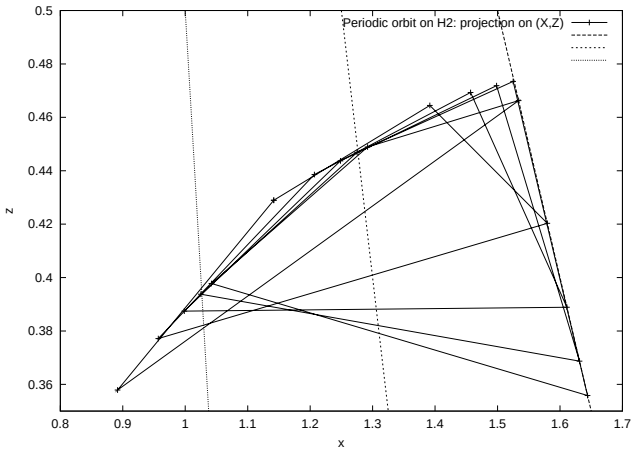

Figure 7: Period of Example 2

Example 3 Consider the three class, two router network of the left part of Figure 2 with $R_x = R_y = R_z = 1$, $p_x = p_y = p_z = 1/2$ but this time with $c_1 = 10^5, c_2 = C$. Figure 8 gives the period of the billiards process as a function of C.

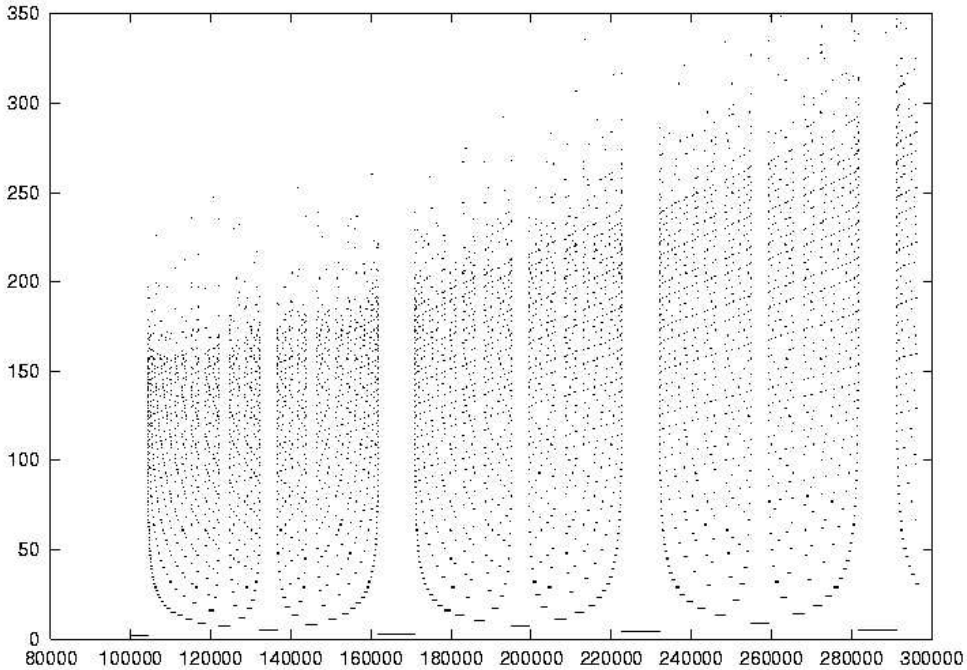

Figure 8: Billiards Period as a Function of C

Figure 8 suggests that the period achieves constant integer values on a Cantor type set of the horizontal axis; in addition, this figure gives numerical evidence that there are infinitely many values of C for which either the right or the left limit of the period is infinite. For instance the period is constant and equal to 2 on the interval $[2, C_0)$ with $C_0 \sim 104295$, whereas the right limit of the period at this point seems to be $+\infty$.

The impact of this phenomenon on average throughput is exemplified on Figures 9 and 10, where we plot the mean throughput w.r.t. C in the neighborhood of C_0. Class 1 takes advantage of the increase of C; there is no such monotonicity for class 2 nor for class 3. Notice the very irregular shape of mean throughput (which is itself a fractal as shown by the zoom) and the singularity at C_2. Notice the similarity with the shape obtained for the same kind of functions from a packet level model of window flow control over a two router network in [3].

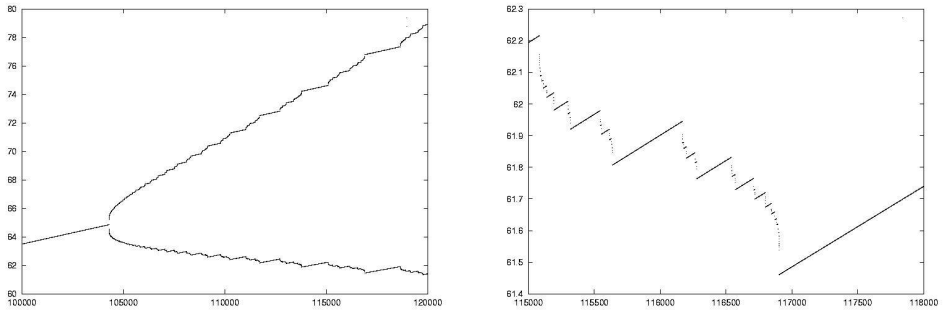

Figure 9: Left: Mean Throughput of Class 1 (upper curve) and Class 2 (lower one) of Example 3. Right: Zoom for Class 2

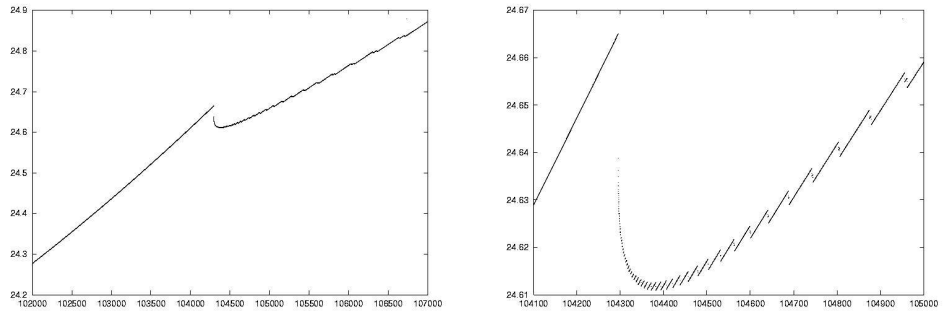

Figure 10: Example 3: Mean Throughput of Class 3 and Zoom

Example 4 This is the 6-class, 3-router network of the right part of Figure 11 also with $R = 1$, $p = 1/2$ for all coordinates. The default value for the speed of a router is $C = 10^5$.

The left part of Figure 12 plots the successive values achieved by the instantaneous throughput of one of the single hop flows of Example 4 versus time (there is here one tick of time when the ball hits one of the billiards facets and we interpolate two successive points via a straight line).

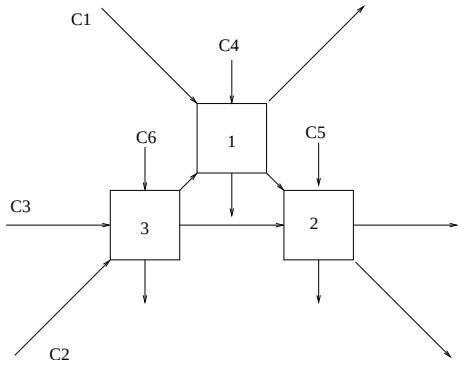

Figure 11: Example 4: Triangle Topology Network

In the right part of Figure 12, we plot the same function under the assumption of some slow trend on the network parameters. Such a trend might stem e.g. from slow changes of the population parameters n_s, which has the same effect as changing the speed parameters as shown by (8).

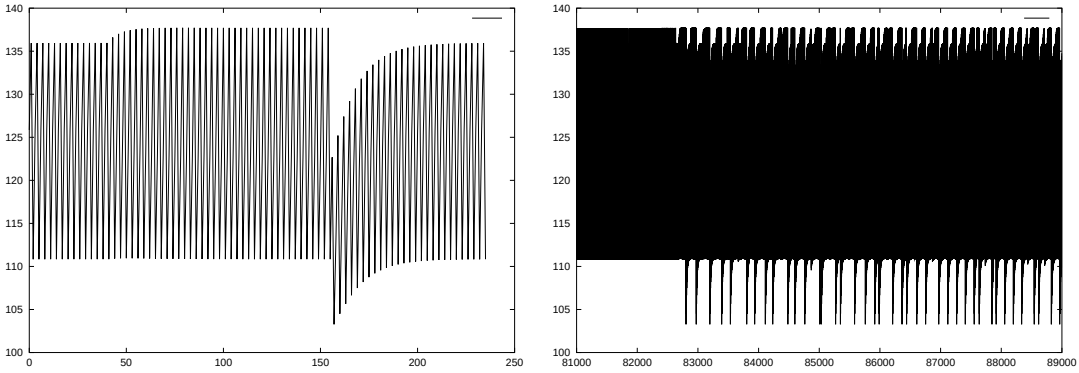

Figure 12: Instantaneous Throughput versus Time under Population Trend for Example 4

Figure 13, where we plot the successive values achieved by this function, without interpolation but under the same trend assumption as above, shows that the set of achieved values lives on a fractal.

3.3 Fairness

3.3.1 Single Router, Several RTTs

Assume that a single router is shared by several classes that only differ through their RTT. Let R_i denote the RTT of class i and p_i its synchronization rate (that we will later take as a function of the average rate). It follows from (7) that a typical flow of class i satisfies the stochastic recurrence: $X_{n+1}^{(i)} = \gamma_{n+1}^{(i)}(X_n^{(i)} + \frac{\overline{\tau}_n}{R_i^2})$ where the sequence $\{\gamma_n^{(i)}\}$ is i.i.d. As a consequence of results in [13], $\overline{\tau}_n$ then converges to a constant $\overline{\tau}$. Taking expectations in the last equation determines the stationary throughput at congestion epochs ($\overline{X}^{(i)}$). Within this setting, the stationary throughput in continuous time is obtained from $\overline{X}^{(i)}$ via the relation $\lambda_i = \overline{X}^{(i)} + \overline{\tau}/2R_i^2$ (see Section 4.1 in [4]). Elementary

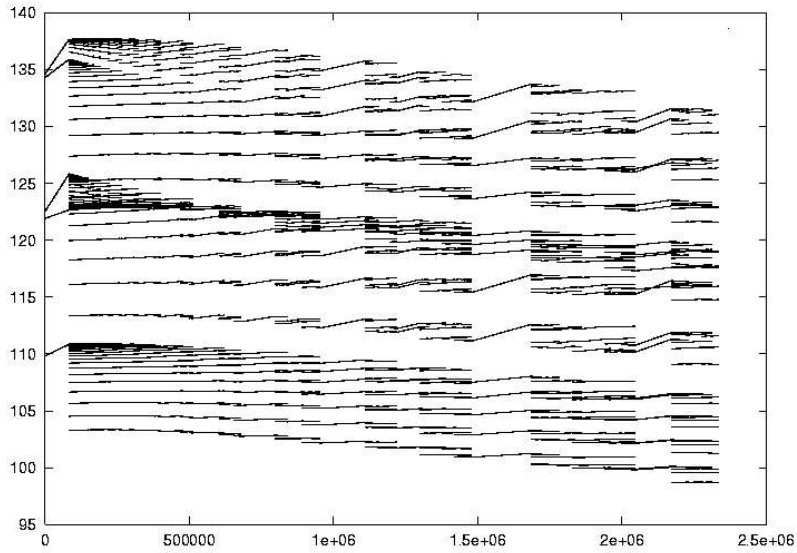

Figure 13: Instantaneous Throughput versus Facet Counter under Population Trend for Example 4

manipulations give:

$$\lambda_i = \left(\frac{\overline{\gamma}^{(i)}}{1-\overline{\gamma}^{(i)}} + \frac{1}{2} \right) \frac{\overline{\tau}}{R_i^2} = \frac{1}{2} \frac{1+\overline{\gamma}^{(i)}}{1-\overline{\gamma}^{(i)}} \frac{\overline{\tau}}{R_i^2} = \frac{4-p_i}{2p_i} \frac{\overline{\tau}}{R_i^2}. \tag{11}$$

So for all i, j, we have:

$$\frac{\lambda_i}{\lambda_j} = \frac{R_j^2}{R_i^2} \frac{(4-p_i)p_j}{(4-p_j)p_i} \sim \frac{R_j^2}{R_i^2} \frac{p_j}{p_i}, \tag{12}$$

where the last equivalence is when the synchronization rates are small. If we assume that synchronization probabilities are proportional to the rate λ_i, i.e., $\frac{p_i}{p_j} = \frac{\lambda_i}{\lambda_j}$, then throughput is proportional to the inverse of RTT (cf. [15, 21, 6, 25]). If we assume that p_i does not depend on the throughputs, we get throughputs proportional to the square of the inverse of RTT. Simulations based on the tools described in §4 confirm this fact (cf. Figure 20).

Let us now concentrate on the RD case. If p_i is of the form

$$p_i = \beta(1 - \exp(-\lambda_i \delta)), \tag{13}$$

as suggested by (19) in Theorem 2, then the stationary throughputs should satisfy the following fixed point equation:

$$\frac{\lambda_i}{\lambda_j} = \frac{R_j^2}{R_i^2} \frac{(1-\exp(-\lambda_j\delta))}{(1-\exp(-\lambda_i\delta))}. \tag{14}$$

If $R_i < R_j$, then $\lambda_i > \lambda_j$, and hence $p_j < p_i$. In addition, from (13), $p_i/p_j < \lambda_i/\lambda_j$. Therefore we always have:

$$\frac{R_j}{R_i} < \frac{\lambda_i}{\lambda_j} < \left(\frac{R_j}{R_i}\right)^2. \tag{15}$$

This confirms experimental studies (cf. [16]) which suggest that the ratio λ_i/λ_j is always proportional to $(R_j/R_i)^a$ with $1 < a < 2$. Let us identify the possible values of a from our analytical framework. When $\lambda\delta$ (defined in (13)) is small, we see that (15) is valid indeed with a close to 1; since δ in (13) is common to all flows, $\lambda\delta$ will be small for the slow flows (here those with large RTTs). Similarly, for those sources with $\lambda_i\delta$ large enough (the fastest flows, or equivalently here those with small RTTs), p_i is close enough to 1, and hence a is close to 2. So, if there is a large enough range of RTT's, the logarithm of the stationary throughput should be a linear function of the logarithm of the RTT, with a slope that is close to -1 for small throughputs, and close to -2 for larger throughputs.

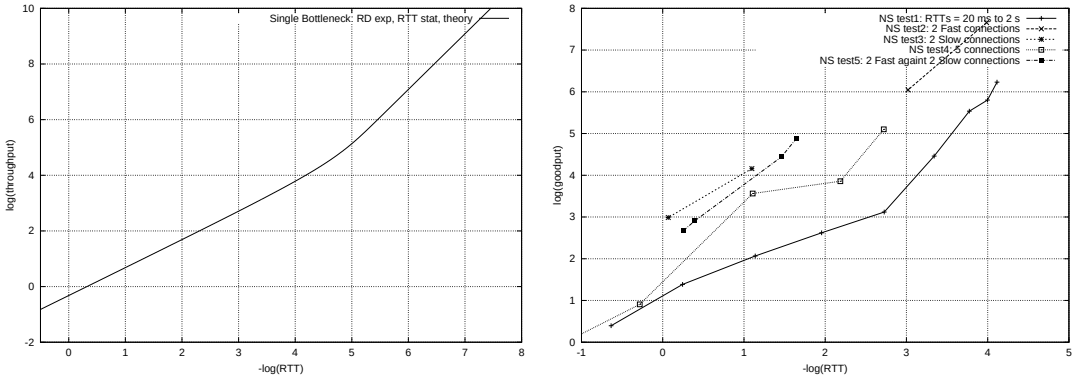

Figure 14: Left: Throughput vs RTTs in log-log as predicted by theory. Right: NS simulation of 5 different cases when varying parameters (number of flows, capacities, propagation delays).

The left part of Figure 14 gives the corresponding log-log plot which is obtained from the model based on (14). In order to solve this fixed point equation, we proceed as follows: for each value of λ, we can find a value of $R = RTT$ satisfying (11) and (13).

It is interesting to see that this predicts a clear breakpoint between the region $a = 1$ and $a = 2$. This is also what we observe in NS simulations under some conditions that correspond to our assumptions: in particular timeouts ought to be negligible and the number of flows of each class should be large. NS simulations of hundreds of different cases with 2 to 10 parallel flows have been studied. A few of them are plotted in Figure 14 (Right). Notice that only the slope of these plots are meaningful. Here are some remarks:

- timeouts when present increase the slope for large RTT flows;
- most of time, one observes a slope between 1 and 2;
- in 10-20% of the cases, we obtained strange results, possibly due to the very deterministic behavior of NS: e.g. $a \sim 4$, even when timeouts were not present, or $a < 0$.

In any case, both our analytical approach and the experimental test mentioned above indicate that as soon as one handles networks with significantly different RTTs (which will be the case for most large networks), representations of TCP that are primarily based on the assumption that throughputs are

linked by an equation of the type $\lambda_i/\lambda_i = R_j^a/R_i^a$ with a constant are probably inadequate for the whole range of values of the RTTs.

Other RD models (such as for instance those obtained when taking Y_n^i in place of λ_i in (13) – see §2.6) also lead to two linearity regions ($a \sim 1$ and $a \sim 2$) with a threshold that separates them in a rather sharp way (see §4).

3.3.2 Two Routers, Several RTTs

Let us revisit Example 1 (§3.1.3) with some more general parameters. The RTT of class i is R_i and its synchronization rate p_i (so here we do not assume that the CI assumption holds anymore). Whenever the sequence of facets is periodic with period two (this is the case when $R_x = R_y$ and $\overline{\gamma}^y = \overline{\gamma}^x$ as already mentioned), one can then identify the periodic regime from the following set of affine equations:

$$\mu_x = (\mu'_x + \frac{\overline{\tau'}}{R_x^2}), \quad \mu_y = \overline{\gamma}^y(\mu'_y + \frac{\overline{\tau'}}{R_y^2}), \quad \mu_z = \overline{\gamma}^z(\mu'_z + \frac{\overline{\tau'}}{R_z^2})$$

with

$$\mu'_x = \overline{\gamma}^x(\mu_x + \frac{\overline{\tau}}{R_x^2}), \quad \mu'_y = (\mu_y + \frac{\overline{\tau}}{R_y^2}), \quad \mu'_z = \overline{\gamma}^z(\mu_z + \frac{\overline{\tau}}{R_z^2}).$$

Direct calculations lead to:

$$\lambda_x = \frac{4-p_x}{2p_x}\frac{T}{R_x^2}, \quad \lambda_y = \frac{4-p_y}{2p_y}\frac{T}{R_y^2},$$
$$\lambda_z = \frac{2}{p_z(4-p_z)}\frac{T}{R_z^2}\left(\frac{8-4p_z+p_z^2}{4} - \frac{T'p_z^2}{2T^2}\right)$$

with $T = \overline{\tau} + \overline{\tau'}$ and $T' = \overline{\tau}\overline{\tau'}$. We see that for the ratio λ_x/λ_y, the result is as the single router case. For the other one, we have:

$$\lambda_z/\lambda_x = \frac{R_x^2}{R_z^2}\frac{4p_x}{p_z(4-p_z)(4-p_x)}\left(\frac{8-4p_z+p_z^2}{4} - \frac{T'p_z^2}{2T^2}\right).$$

Hence

$$\frac{\lambda_z}{\lambda_x} = \frac{R_x^2}{R_z^2}\frac{p_x(4-p_z)}{p_z(4-p_x)} \times \alpha$$

with $1/3 \leq \alpha \leq 1/2$. This means that even if the flow that crosses the two routers had the same RTT as the two others, in the best case (p proportional to λ) this flow is 30% slower than the two others; in the worst case ($p_x = p_y = p_z$), it could be 3 times slower than the others. Now if its RTT is twice larger than that of the other flows, then the best it can get is 3 times slower compared to the others, and in the worst case its connection is 12 times slower!

3.3.3 Fairness in the Non-Periodic Case

The aim of this section is to analyze bandwidth sharing as a function of the network parameters, and in particular the speed of the routers.

The following two figures illustrate bandwidth sharing for Example 4. We plot the throughput obtained by certain classes against that obtained by other classes, when varying the speed C_1 of router 1 on some interval. We do this both for mean throughput and for instantaneous throughput (the set of values actually achieved by the throughput process when sampled at the hitting times of a certain face).

For $C_1 = 84000$, we get approximately $\lambda_1 = \lambda_2 = 33.8$, $\lambda_3 = 42.5$, $\lambda_4 = 132.4$ and $\lambda_5 = \lambda_6 = 99.5$. In Figure 15, we plot the sum of the mean throughput of all classes that use router 3 (classes 2,3 and 6) against the mean throughput of the 2-hop class that does not use router 3 (class 1) for C_1 Ranging from 7300 to 8700 Approximately. We again observe a fractal and a non-monotonic behavior.

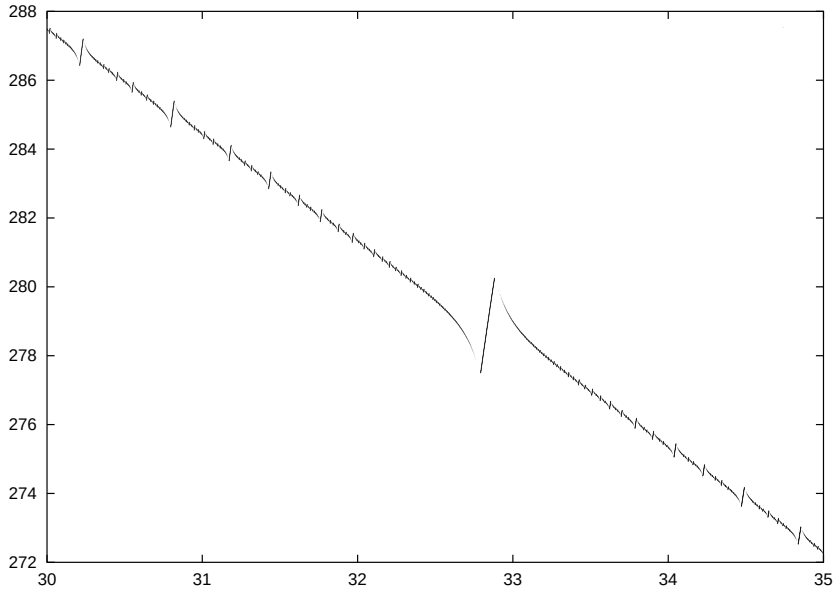

Figure 15: Example 4: Sum of the Mean Throughputs of Classes 2,3 and 6 w.r.t. that of Class 1

The upper part Figure 16 plots the instantaneous throughput of classes 1 and 6 w.r.t. the instantaneous throughput obtained by class 2 (these three classes are those sharing router 1). The lower part is a zoom of the latter.

Here also, we find a general trend, but a quite complex fractal behavior along this trend, which leaves little hope for simple closed form formulas.

When playing with parameters, such fractals show up in all topologies (not reduced to one router) with a wide variety of shapes. A collection of fractals generated by this class of interaction models can be downloaded at http://www.di.ens.fr/∼trec/aimd

3.4 A Conservation Law

Assume a stochastic TCP billiards admits a stationary regime. Assume in addition that its synchronization rate is rate and class-independent (see §2).

Let ν_r denote the (continuous time) intensity of the congestion epochs of router r. Let $S(t) = \sum_{i,s} X^{(s,i)}(t)$ be the sum of all flow throughputs in continuous time.

- But for a denumerable set of discontinuities, $S(t)$ is linearly increasing with the rate $\sum_s N_s/R_s^2$.

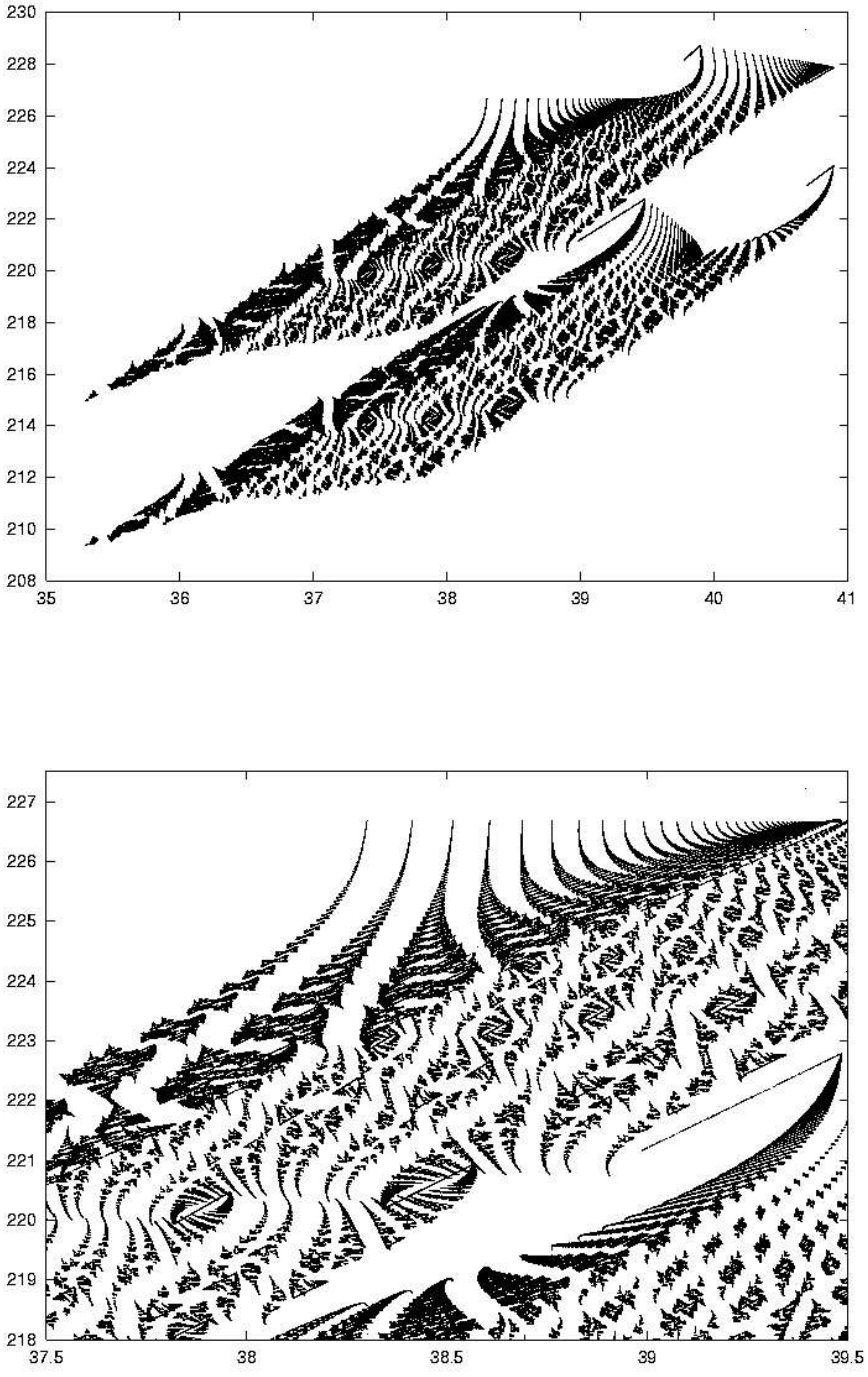

Figure 16: Instantaneous Bandwidth Sharing for Example 4 and Zoom

- Because of the class-independent assumption, each type r congestion epoch creates a jump of $S(t)$ downward of mean magnitude $C_r(1-\overline{\gamma}^r)$.

The drift upward should compensate the jumps downward, so that the following conservation law necessarily holds:

$$\sum_{s\in\mathcal{S}} \frac{N_s}{R_s^2} = \sum_{r\in\mathcal{R}} \nu_r C_r (1-\overline{\gamma}^r). \tag{16}$$

This readily implies the following relation for the associated deterministic Billiards (where ν_r has the same interpretation as above):

$$\sum_{r\in\mathcal{R}} \sum_{s\in\mathcal{S}_r} \frac{a_{s,r}}{R_s^2} = \sum_{r\in\mathcal{R}} \nu_r c_r (1-\overline{\gamma}^r). \tag{17}$$

3.5 Implications and Connections

The practical implications of the long periodic or non-periodic behavior illustrated in §3.2 and 3.3.3 should be understood when taking into account that these results only hold for the large population asymptotic model with $N=\infty$. To capture the behavior of any model with finite population, one should of course add small Gaussian fluctuations to this, which will result into a blur of the overall dynamics and of the limiting sets describing bandwidth sharing and throughput.

Nevertheless, these results suggest that for large populations and non-zero synchronization, *aggregated throughput* (that is the empirical mean process) exhibits a rather complex behavior which leads to fluctuations that are due to the network as a whole, and are to be added to other and more classical flow or packet level fluctuations.

The properties reported in these two subsections are different from the simulation based observations on the chaotic behavior of TCP reported in [27]. The main difference lies in the fact that the properties of the present paper bear on the sensitivity of aggregated traffic (and more precisely on the empirical mean values as defined above) w.r.t. some topology parameter (e.g. the speed of a router) whereas the observations of [27] focus on the dependence of the throughput of individual flows w.r.t. initial conditions for a given topology.

There might however be a link between the sensitivity w.r.t initial conditions and the fact that the facet and billiards could have a non-periodic behavior for a given topology.

4 Simulation

The stochastic Multi-AIMD equations (5) are the basis of the simulator used in this section. Like in discrete event simulation, these equations are efficient in that they allow one to jump from a congestion epoch to the next. A slotted version of the non-linear AIMD model (§2.6) was also implemented; this version, which is the one used in the results reported on below, allows one to take the most important missing effects into account, including, buffer contents tracking, the effect of buffer size on RTT, the slow start etc. The aim of this section is twofold.

- We first show that this simulator allows one to study fine properties of quite large networks, including the sensitivity of throughput w.r.t various network parameters. This is possible because the simulation cost is approximately linear in the number of the congestion epochs and also in the number of TCP flows.
- We then study the statistical properties of the aggregated traffic generated by this model and compare them to what is reported in the literature.

4.1 Network Topology

The network topology that is studied is featured on Figure 17. In addition to the hierarchical traffic (which will be referred to as the main traffic below), we often add some cross-traffic flows to routers of certain levels. By definition, a cross traffic flow uses this specific router only.

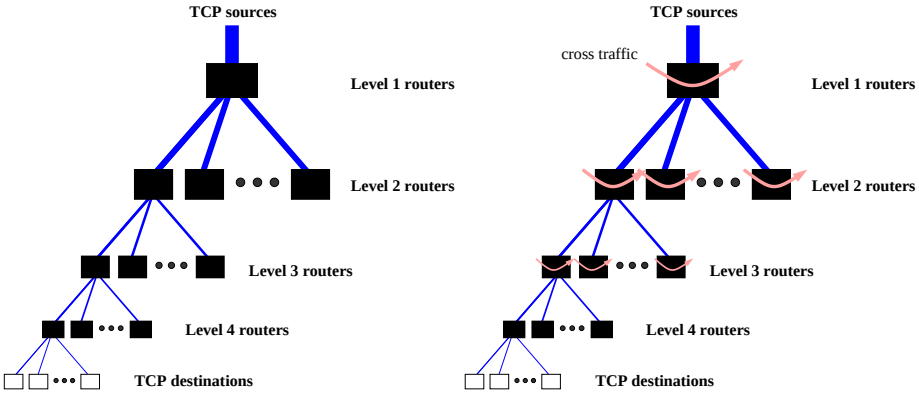

Figure 17: Tree Topology. Left: without cross traffic. Right: with cross traffic.

4.2 Bottleneck Analysis

The simulator shows that for most configurations, and in particular for configurations as those depicted in Figure 17, there are more than one bottleneck router (or here bottleneck level) for a given flow. So, the relevant variables are in fact the proportions of congestion epochs (bottlenecks), and the proportion of losses (MD's), that are of a given type (or level) over time in the stationary regime. The two should be distinguished because of the synchronization rates: a bottleneck or congestion epoch at level 1 might create a huge number or losses or MD's even with a moderate synchronization rate, which is not the case at higher levels.

In some particular cases, such as the case where all RTTs are exactly the same and where in addition, there is no cross traffic, a single router approximation could possibly be used. But even in this case, when varying capacities, the transition of the bottleneck from one level to another is not instantaneous. The stationary proportions in question are plotted for this case on the left part of Figure 18, which shows the variations of these proportions when increasing the service capacity of the level 2 routers (which is the bottleneck level on the left part of the plot). In this case, the network is a three level tree. Each router of level 3 is an access router with 10 long lived TCP flows and has a capacity C_3. Level 2 routers are concentrating the flows from 20 routers of level 3 and have a capacity of $C_2 = 10$ Mb/s. The level 1 router concentrates the flows of 30 routers of level 2 and has a capacity of $C_1 = 300$ Mb/s. When $C_3 = 450$ Kb/s, the leaves of the tree are the bottleneck. The transition of the MD proportion curves is rather fast: with a variation of 1.4% of C_3, the MD proportion varies from 100%-0% to 50%-50%. The transition of the bottleneck curves requires an increase of 20 % of the initial bottleneck capacity C_3 (from the value 500 Kb/s).

This implies that aggregated traffic seen from Level 2 has statistical properties that are very sensitive w.r.t. capacity characteristic (cf. the right curve of Figure 18).

Figure 19 provides a very simple example of situation where the synchronization rate could introduce some unexpected throughput behavior. In this case, we have a single bottleneck router r shared

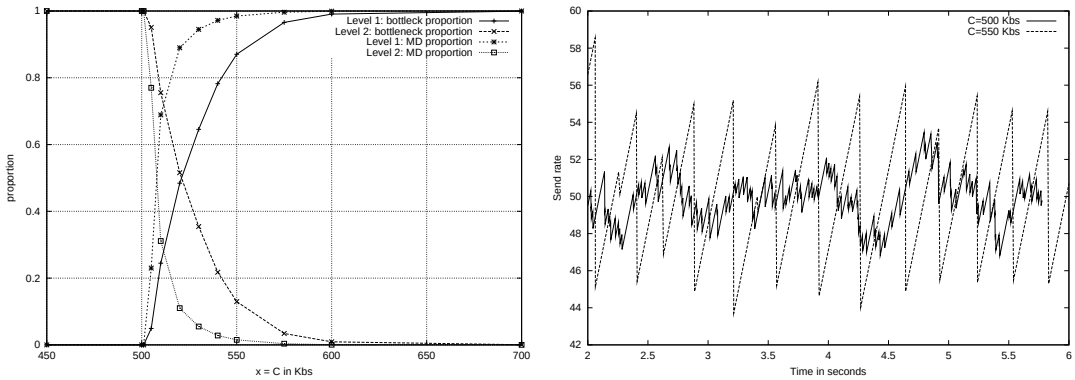

Figure 18: Bottleneck Transition. Left: loss and bottleneck proportions; Right: Level 2 aggregation of TCP flows.

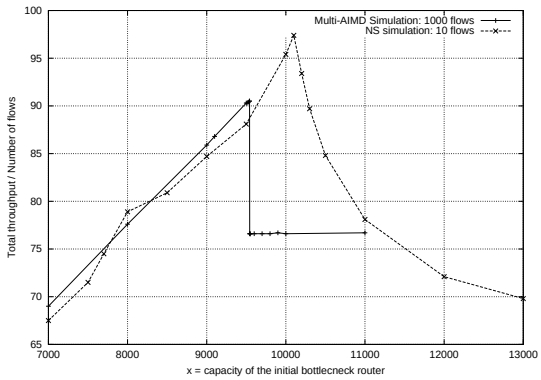

Figure 19: Non-monotonicity due to synchronization effect.

by N flows and we choose its buffer size and the RTTs of the flows such that the synchronization rate at this router is small. We then connect router r to a second router r' (on the common route of all flows) with a very small buffer size (so that the synchronization at this router should be important) and with a speed 30% bigger than that of R_2. Keeping everything unchanged, if one increases the speed of router r, one should observe a transition of the bottleneck and when r' becomes bottleneck; surprisingly enough, what we observe when doing so is actually a throughput drop. This is actually observed both on a NS simulation and on the AIMD simulation. In the NS simulation we checked that the drop is not due to timeouts but actually to the high synchronization.

4.3 Sensitivity w.r.t. RTT's

Consider a two level hierarchical network within the class described above, with bottlenecks at two different levels: (local) bottleneck routers, located at the lower level, each of which concerns a group of 100 flows, and a global bottleneck at the higher lever, which concentrates 50 lower level routers, that is 5000 flows. This time however, RTTs are heterogeneous (sampled uniformly from 1 ms to 2 s). The analysis of the proportions of losses per class obtained from the simulator shows that slow flows are less affected by the local bottleneck, whereas the fast ones are mainly affected by it. Nevertheless, the simulator shows that as in the single router case, within a group of fast flows, we still have $a \sim 2$,

whereas a group of slow flows, $a \sim 1$. In this multi router case, the transition between these two groups seems to be more progressive than in the single bottleneck case.

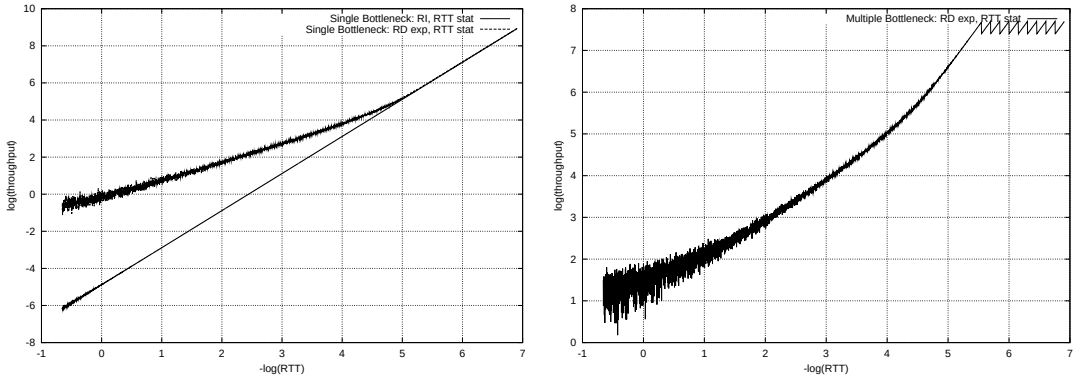

Figure 20: Throughput vs RTTs in log-log. Left: single router, comparison of RI and RD; Right: Multi router case, several bottlenecks at different levels.

4.4 Aggregated Traffic Analysis

In this section, the network is a 4 level tree. The number of routers of level $n+1$ is 10 times the number of routers of level n. In Case 1, the routers capacities are equal to 500 Mb/s, 50 Mb/s, 5 Mb/s, 500 Kb/s, and the buffer capacities are equal to 10000, 1000, 100, 10 packets respectively. For Case 2, the capacity of the last level (leaf) routers is increased to 1 Mb/s and the buffer capacities are modified to 6000, 2000, 100, 10 respectively. We generated 4 classes of propagation delays (RTTmin): 0.1, 0.2,0.3,0.4 seconds with equal probability, the mean RTT seen by one flow being then approximately equal to RTTmin plus approximately 0.35-0.45 s.

The simulation results for Case 1 are given on the top of Figure 21. The leftmost curve concerns the case when losses are at the leaves of the tree (all of them in this case). When aggregating a larger number of flows, the fluctuations decrease as predicted by the law of large number. The rightmost curve gives both the transient and the stationary parts of the aggregated throughputs. The convergence to stationary regime is most of time exponentially fast [5].

The bottom part of Figure 21 features Case 2. In this case, losses are present at all levels of the tree and fluctuations are not erased by aggregation, even if level 1 is very rarely bottleneck. In this case, 80.4% of bottlenecks are at level 4, 17.4% at level 3, 2% at level 2 and 0.2% at level 1. The respective mean synchronization rates are 0.50, 0.39, 0.28 and 0.22. The mean throughput averaged over the 4 classes is 35 Kb/s. The value of C_r/N_r (router capacity divided by the number of flows sharing this router) is 45, 45, 45 and 90 Kb/s for levels 1,2,3 and 4 respectively. Therefore the global under-utilization is of 22% ($(1-35/45)\times 100$). This is much more than what the single router AIMD model would predict for the value of the synchronization of the 4 levels: using the formula derived in [4], we would get the following values: $p/4 \times 100 = 12.5, 10, 7, 5.5\%$ for the various levels.

Figure 22 studies that case when the topology of the network is still a 4 level tree, but this time with an additional cross traffic (also made of long lived TCP flows) that is local to each router. This additional traffic is present on routers of all levels (but the highest), and consists of an additional

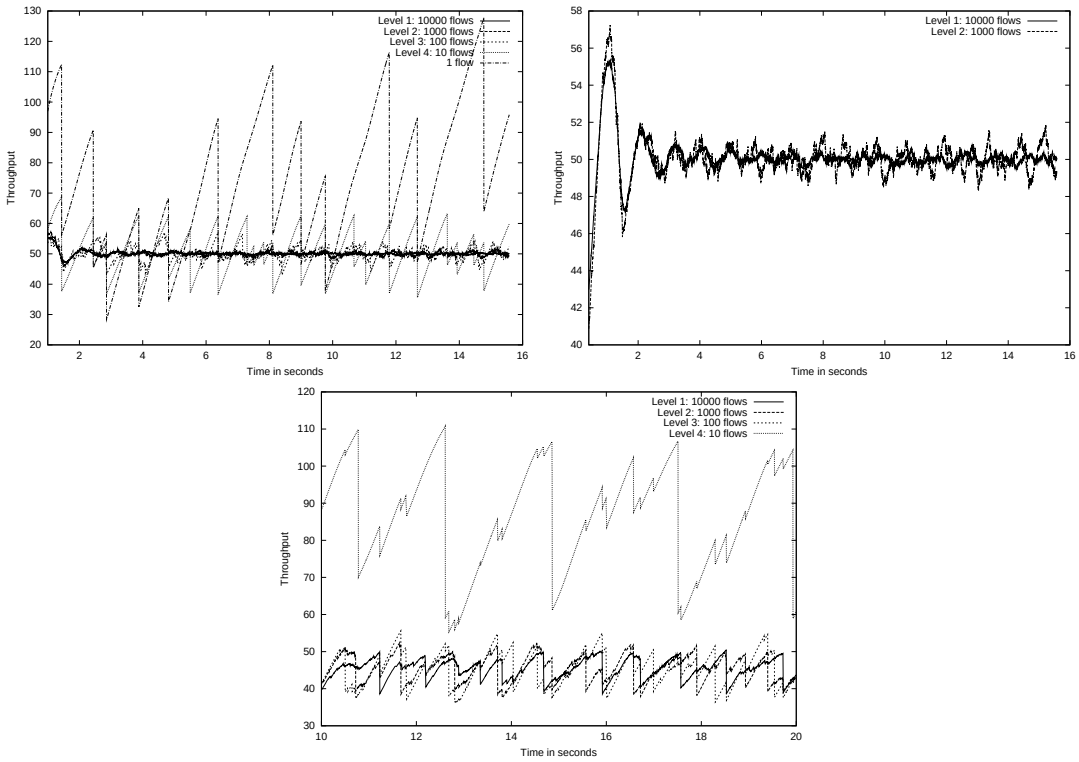

Figure 21: Throughput evolution over time. Left: Case 1, the 5 levels of aggregation (from 1 to 10000 flows). Right: Case 1, the geometric convergence of aggregated traffic to its stationary regime. Bottom: Case 2, aggregation at different levels of the tree.

number of users that amounts to 10% of the main traffic going through this router. In this case, the total number of TCP flows is 1210000 and there are 10211 routers.The routers characteristics are then as follows:

- Level 4: $C = 6Mb/s$, $B = 100$ pkts shared by 100 flows of the main traffic;
- Level 3: $C = 250Mb/s$, $B = 5000$ pkts shared by $100 \times 50 = 5000$ flows of the main traffic and 500 cross flows;
- Level 2: $C = 5Gb/s$, $B = 100000$ pkts shared by $100 \times 50 \times 20 = 100000$ flows of the main traffic and 10000 cross flows;
- Level 1: $C = 50Gb/s$, $B = 1000000$ pkts shared by $100 \times 50 \times 20 \times 10 = 1000000$ flows of the main traffic and 100000 cross flows.

In this case, losses necessarily occur at each level due to the presence of the cross traffic. Taking capacities (buffer and speed) of each leve proportional to the number of flows at this level leads to a synchronization rate that does not vary too much from level to level: 0.48, 0.46, 0.36, 0.45. As it was already noticed in Case 1, the higher levels of the tree are less often bottleneck: we get here 0.02, 0.19, 4.4 and 95 % from top to bottom.

In this simulation, we first remark that when aggregating flows we don't always have a significant decrease of the fluctuations. When losses occur at the top level with a given synchronization rate p,

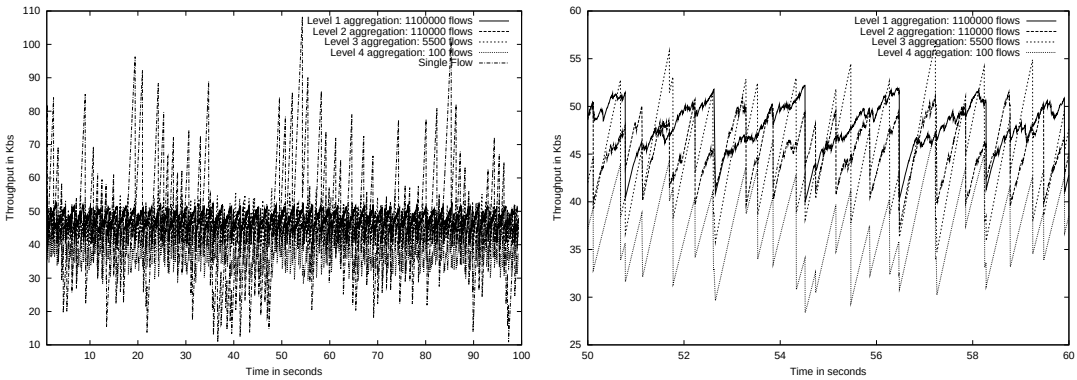

Figure 22: Throughput evolution over time. Left: the 5 aggregation levels (from 1 to 1000000 flows); Right: Zoom on 4 levels.

this level aggregation and all its children level should have a fluctuation of at least $100 \times p/2\%$. In this case, we observe $p = 0.246, 0.257, 0.382, 0.741$ respectively from top to bottom level and more of 90% of the bottlenecks were at level 3.

We tested the statistical properties of traffic aggregated at different levels using the Matlab tool developed by P. Abry and D. Veitch [1]. Below, we plot the second order logscale diagram (LD) of the energy function and the multiscale analysis (MS) diagram (for more on these questions, see e.g. [2]). We see that the different levels exhibit statistical properties for LD and MS plots which are similar to those observed in [4] for the single router case, and which are compatible with a multi-fractal scaling.

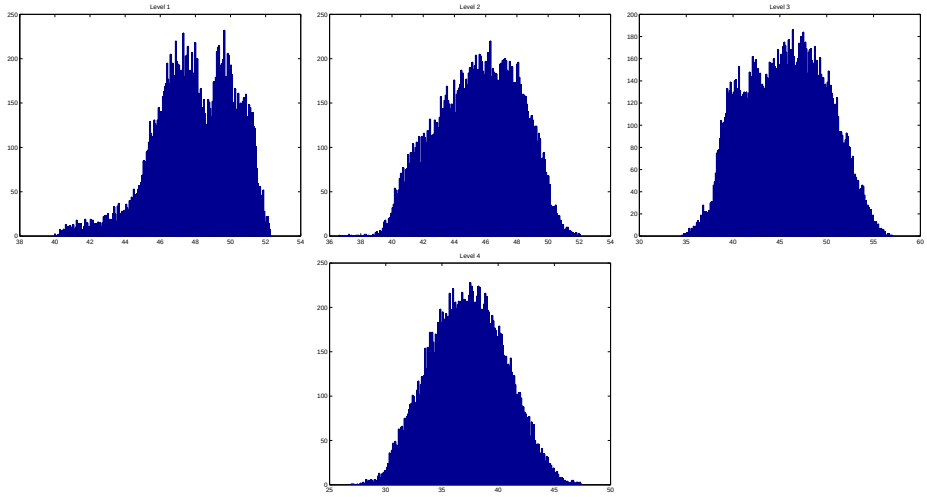

Figure 23: Empirical Distribution Function for aggregated traffic at different levels (from level 1 left to level 4 right)

It is interesting to observe that whereas the empirical distribution functions we obtained exhibit quite different shapes (see e.g. the level 1 with two peaks compared to the more Gaussian like distribution of level 4), the LD and MS analysis are quite insensitive w.r.t. the level of aggregation. The scaling exponent α is between 1.83 and 1.96 in all cases.

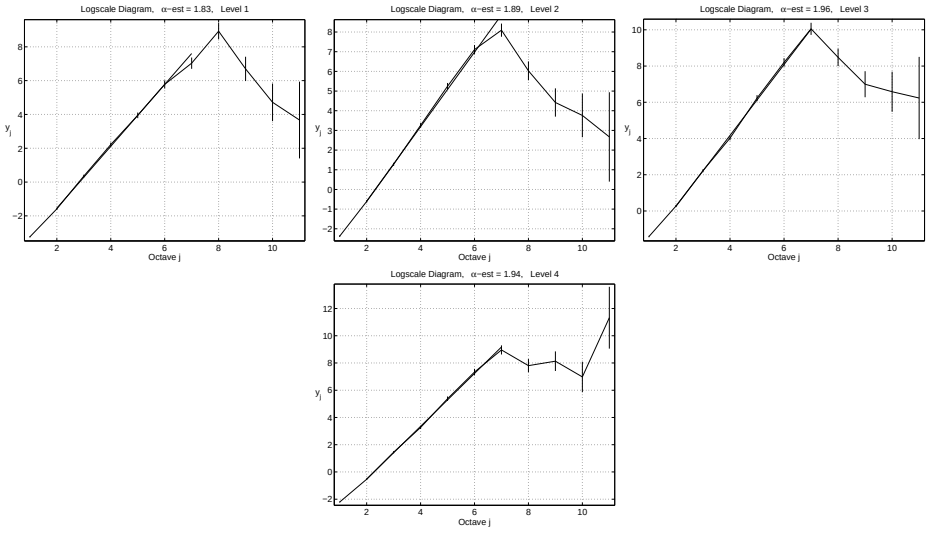

Figure 24: LD plots for different levels (from level 1 left to level 4 right)

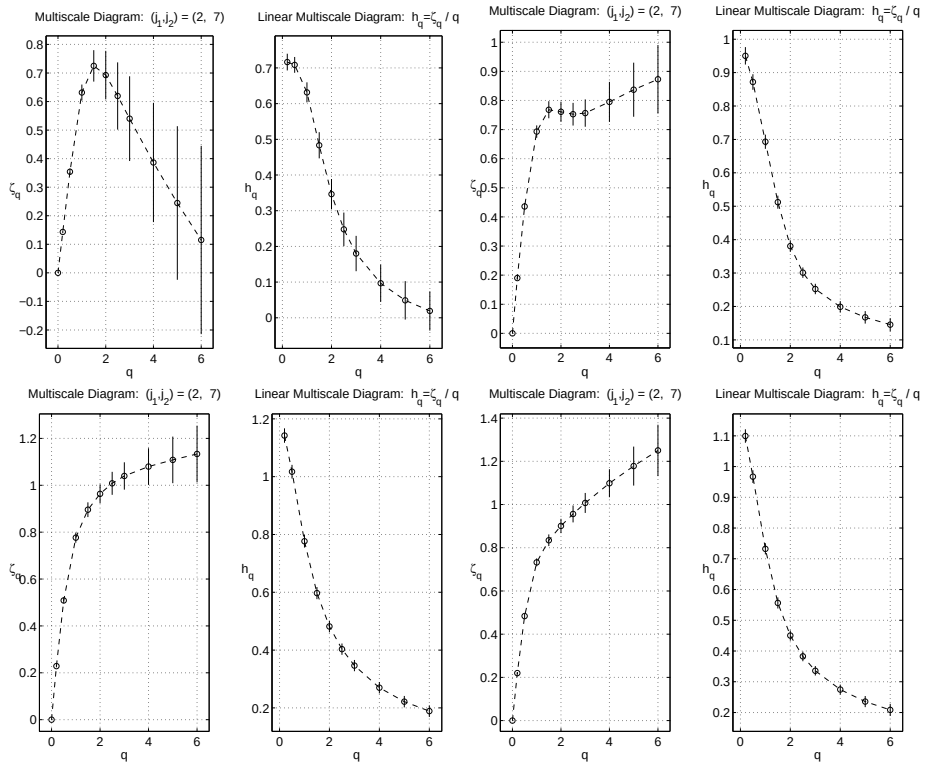

Figure 25: Multiscale analysis for different levels (from level 1 left to level 4 right)

5 Appendix

5.1 Estimation of the Synchronization Rate

5.1.1 Estimation of loss probability at congestion times

In order to estimate the probability that a packet of class s is lost at a congestion epoch of type r, with $r \in \mathbb{P}_s$, we use a simple $M/M/1/B_r$ approximation. The argument goes as follows. During

the congestion period, the total arrival rate is approximately equal to the total service rate, namely C_r. This lasts for a duration of approximately $\eta_r = \min_{s \in \mathcal{S}_r} R_s$, since the flows with the shortest RTT then react to losses on router r, thus ending the congestion period. For the (multiclass) $M/M/1/B_r$ queue with total arrival and service rates such that $\lambda = \mu$, the steady state packet loss probability is $L_r = 1/(B_r + 1)$ regardless of the class. If $B_r/C_r << \eta_r$, it makes sense to approximate the empirical frequency with which packets of any type s, with $r \in \mathbb{P}_s$, are lost on $[0, \eta_r]$ by this stationary probability. In a refined model, we use the same argument but with $\lambda = C'_r$ and $\mu = C_r$, with for instance C'_r the arrival rate at time η_r if the buffer is empty at time 0, the total arrival rate is C_r at this time, and all flows sharing r increase their throughput according to the AI rule on the interval $[0, \eta_r]$, that is

$$C'_r = C_r + \eta_r (\sum_{s \in \mathcal{S}_r} N_s / R_s^2),$$

or a similar formula with $C_+ n^r$ in place of C_r if we take the buffer into account as suggested above. This leads to the following formulas :

$$\underline{L}_r = \frac{1}{B_r + 1}, \quad L_r = \frac{\rho_r^{B_r}(\rho_r - 1)}{\rho_r^{B_r+1} - 1}, \quad \overline{L}_r = \frac{\rho_r^{B_r}}{B_r + 1} \tag{18}$$

where $\rho_r = C'_r/C_r$. We have $L_r \sim \overline{L}_r$ when $\rho_r \to 1$ and $\underline{L}_r = L_r = \overline{L}_r$ when $\rho_r = 1$.

5.1.2 Estimation of the synchronization rate

We now propose an estimation of the synchronization rate also based on the $M/M/1/B_r$ queue analysis and when assuming that $B_r/C_r << \eta_r$. In order to compute the synchronization rate of type r for flows of class s, with $r \in \mathbb{P}_s$, we have to evaluate how many *flows* of this class experience loss during this congestion period, while taking into account the fact that if a given flow has already experienced a loss, then any further packet loss of this flow that takes place in the very same congestion period should not be counted as a new flow loss. At the beginning of the n-th congestion period, the loss rate of flows of type s coincides with the packet loss rate for this class and is equal to $\sum_{i \in s} Y_n^{(s,i)} L_r$. In the asymptotic model, this is close to $N_s y_n^{(s)} L_r$. After the first loss of this class took place, the flow loss rate becomes $(N_s - 1) y_n^{(s)} L_r$; after the second flow loss, the flow loss rate is $(N_s - 2) y_n^{(s)} L_r$, etc. In order to determine the mean number of flows of class s that experience at least one loss by time η_r, we have to study the transient mean value of the continuous time Markov chain (pure birth process) on the integers with transition rates $\lambda_{i,i+1} = (N_s - i) y_n^{(s)} L_r$. Doing so, we obtain :

Theorem 2 *In the large population asymptotic model, from the $M/M/1/B_r$ queue model described above, the synchronization rate for the flows of class s at time T_n is $p_n^{(s,\overline{r}_n)}$ with the function $p_n^{(s,r)}$ given by the formula*

$$p_n^{(s,r)} = \frac{1 - e^{-y_n^{(s)} \eta_r L_r}}{1 - e^{-C_r \eta_r L_r}}. \tag{19}$$

Proof Let $N_s(t)$ be the pure birth process described above. It admits the stochastic intensity $\lambda(t) = y_n^{(s)} L_r(N_s - N_s(t))$. So, from the stochastic integration formula,

$$\begin{aligned} E[N_s(t)] &= E\left[\int_0^t N_s(du)\right] \\ &= E\left[\int_0^t y_n^{(s)} L_r(N_s - N_s(u))du\right] \\ &= y_n^{(s)} L_r N_s t - y_n^{(s)} L_r \int_0^t E[N_s(u)]du. \end{aligned}$$

So, $g(t) = E[N_s(t)]$ satisfies the differential equation

$$g'(t) = y_n^{(s)} L_r N - y_n^{(s)} L_r g(t),$$

with initial condition $g(0) = 0$. The solution is $g(t) = N_s(1 - \exp(-y_n^{(s)} L_r t))$.

By the same argument, the expected number of flows that experience at least a loss given that this number is positive is $h(t) = N_s(1 - \exp(-y_n^s L_r t))/(1 - \exp(-C_r L_r t))$.

So the proportion of flows of class s that experience at least one loss by time η_r given that at least one flow looses is as given in the theorem. □

In order to test the accuracy of Formula (19), we compared the performance result with NS simulations. For a single router bottleneck case with tens of parallel sessions, NS simulations give performance results that have a variation of 10 to 20% when changing parameters others than the capacities and propagation delays. This variation seems to decrease when the number of flows increases, and also when timeouts are negligible. Our performance prediction using the formula (19) is within the range of results given by NS.

5.2 Proof of Theorem 1

The proof is by induction: when summing up (5) over all $i \in s[N]$ and dividing by N_s, when N tends to ∞, we get $x_{n+1}^{(s)}$ on the left hand side provided the right hand side also converges to a limit. On the right hand side, when making use of the induction assumption, we obtain that the argmin of the random variable

$$\min_{r \in \mathcal{R}} \frac{C_r - \sum_{j,u \in \mathcal{S}_r} X_n^{(u,j)}}{\sum_{u \in \mathcal{S}_r} \frac{N_u}{R_u^2}}$$

tends to a deterministic variable limit which is precisely $\overline{r}_{n+1}$, when N grows large. Due to this and the ergodicity assumption on the random variables $\gamma_{n+1}^{(s,i,r)}$, for fixed r and s,

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N_s} \sum_{i \in \mathcal{S}} \gamma_{n+1}^{(s,i,r_{n+1})} &= \lim_{N\to\infty} \frac{1}{N_s} \sum_{i \in \mathcal{S}} \gamma_{n+1}^{(s,i,\overline{r}_{n+1})} \\ &= \mathbb{E}[\gamma_{n+1}^{(s,i,\overline{r}_{n+1})}] = \overline{\gamma}_{n+1}^{(s,i,\overline{r}_{n+1})}. \end{aligned}$$

Finally, using the induction assumption,

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N_s} \sum_{i \in s} \gamma_{n+1}^{(s,i,r_{n+1})} X_n^{(s,i)} &= \lim_{N\to\infty} \frac{1}{N_s} \sum_{i \in s} \gamma_{n+1}^{(s,i,\overline{r}_{n+1})} X_n^{(s,i)} \\ = \lim_{N\to\infty} \frac{1}{N_s} \left(\sum_{i \in s^+(n+1)} X_n^{(s,i)} + \frac{1}{2} \sum_{i \in s^-(n+1)} X_n^{(s,i)} \right), \end{aligned}$$

where $s^+(n+1)$ is the set of flows $i \in s$ such that $\gamma_{n+1}^{(s,i,\overline{r}_{n+1})} = 1$, and $s^-(n+1)$ the complementary set. From the independence, the induction assumption and the ergodicity assumption,

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N_s} \sum_{i\in s^+(n+1)} X_n^{(s,i)} \\
&\quad = \lim_{N\to\infty} \frac{|s^+(n+1)|}{N_s} \frac{1}{|s^+(n+1)|} \sum_{i\in s^+(n+1)} X_n^{(s,i)} \\
&\quad = (1 - p_{n+1}^{(s,\overline{r}_{n+1})}) x_n^{(s)},
\end{aligned}$$

with a similar argument for the other sum. One proves in the same way that for any subset $\sigma[N]$ of $s[N]$ with a cardinal that grows to infinity with n,

$$\exists \lim_{N\to\infty} \frac{1}{|\sigma[N]|} \sum_{i\in\sigma[N]} X_{n+1}^{(s,i)}[N] = x_{n+1}^{(s)} \quad a.s.$$

so that we can propagate the last property as our induction assumption. The proof of the last property is also by induction. It is quite similar to the one given above and it is omitted here.

5.3 Proof of Lemma 4

Let $\mathcal{P}$ denote the hyperplane $\sum_s x_s = 0$ of $\mathbb{R}^K$. Let $\Pi(x)$ denote the projection of $x \in \mathbb{R}^K$ on this hyperplane along the direction of the vector $\rho = (1/R_1^2, \ldots, 1/R_K^2)$, that is $(\Pi(x))_s = x_s - h_s$ with $h_s = \frac{1}{R_s^2}(\sum_u x_u)$. Let D denote the following projective distance between two vectors x and x' of $\mathbb{R}^K$:

$$D(x,x') = \sum_{s=1,\ldots,K} R_s^2 (\Pi(x))_s - \Pi(x'))_s)^2. \tag{20}$$

Two vectors have a nul D distance if and only if their projections on $\mathcal{P}$ coincide.

Let r be a fixed router. Consider two vectors x and x' that belong to H_r, namely which are such that $\sum_{s\in\mathcal{S}_r} x_s = \sum_{s\in\mathcal{S}_r} x'_s = c_r$. The dynamics from facet H_r to the next facet transforms these vectors as follows:

- They are first transformed into $\widetilde{x}$ and $\widetilde{x}'$ respectively with $\widetilde{x}_s = x_s$ for $s \notin \mathcal{S}_r$ and $\widetilde{x}_s = \overline{\gamma}^{(r)} x_s$ for $s \in \mathcal{S}_r$ (and similar definitions for x'). Here, we used the (CI) assumption which implies that for all r, the synchronization rate on router r is the same for all flows in $\mathcal{S}_r$, that is $\overline{\gamma}^{(s,r)} = \overline{\gamma}^{(r)}$ for all $s \in \mathcal{S}_r$.
- The vectors $\widetilde{x}$ and $\widetilde{x}'$ are then projected onto some hyperplane H_t (the next facet) along the direction ρ, which gives $\widehat{x}$ and $\widehat{x}'$ respectively.

Let us check that when applying both transformations, the D distance between the two vectors cannot increase.

For the first transformation, we have to check that (with the notation introduced for the Π projection)

$$\sum_s R_s^2 (x_s - x'_s - (h_s - h'_s))^2 \geq \sum_s R_s^2 (\widetilde{x}_s - \widetilde{x}'_s - (\widetilde{h}_s - \widetilde{h}'_s))^2.$$

For all $x \in \mathbb{R}^K$, let

$$h(x) = \frac{\sum_{u\in\mathcal{S}} x_u}{\sum_{u\in\mathcal{S}} 1/R_u^2}.$$

Using the fact that for all s, $h_s - h'_s = \widetilde{h}_s - \widetilde{h}'_s = \frac{1}{R_s^2}(h(x) - h(x'))$ and the fact that $(h - h')$ only depends on the coordinates $x_s, s \notin \mathcal{S}_r$, we see that for all $s \notin \mathcal{S}_r$, $(x_s - x'_s - (h_s - h'_s)) = (\widetilde{x}_s - \widetilde{x}'_s - (\widetilde{h}_s - \widetilde{h}'_s))$. So it is enough to check that

$$\sum_{s\in\mathcal{S}_r} R_s^2(x_s - x'_s - (h_s - h'_s))^2 \geq \sum_{s\in\mathcal{S}_r} R_s^2(\overline{\gamma}^{(r)}(x_s - x'_s) - (h_s - h'_s))^2$$

which is immediate since $\overline{\gamma}^{(r)} < 1$ and since

$$\sum_{s\in\mathcal{S}_r} R_s^2(h_s - h'_s)(x_s - x'_s) = \kappa \sum_{s\in\mathcal{S}_r} (x_s - x'_s) = 0,$$

where κ is some constant.

Now $\widetilde{x}$ and its projection $\widehat{x}$ on H_t along the direction of ρ obviously have the same Π projection. So the second transformation preserves D.

We conclude that the last map is non-expansive for the projective distance D. More precisely, the distance decreases of exactly

$$(1 - (\overline{\gamma}^{(r)})^2) \left(\sum_{s\in\mathcal{S}_r} R_s^2(x_s - x'_s)^2 \right),$$

so that in this case, the only possibility for having no strict decrease of the distance is that $x_s = x'_s$ for all $s \in \mathcal{S}_r$.

Consider now a period of the discrete dynamics, say with sequence of facets $r_1, r_2, \ldots, r_n$. Assume that there exist two points $x = x_1$ and $x' = x'_1$ of H_{r_1} such that the D-distance between the orbits $x_1, \ldots, x_n$ and $x'_1, \ldots, x'_n$ of these two points remains constant over the period. What precedes implies that $(x_1)_s = (x'_1)_s$ for all $s \in \mathcal{S}_{r_1}$ and this equality is then preserved over the whole period since the same facets are used. More generally, for all $1 \leq q \leq p \leq n$, $(x_p)_s = (x'_p)_s$, for all $s \in \mathcal{S}_{r_q}$ and $p \geq q$.

Since each facet is visited at least once over the period (thanks to the assumption that each router has at least one flow that uses this router alone), then necessarily $x_n = x'_n$.

6 Conclusion

We have introduced a model allowing one to study the bandwidth sharing operated by TCP on networks composed of several tail-drop routers or links under the assumption of non-zero source synchronization.

This model is based on the interplay between three (sub) models: a deterministic network level model (the billiards), a set of more or less independent stochastic models for individual flows, where the influence of the whole network is taken into account via certain averages, and a packet level model that is only used for determining the delay of reaction of sources and the associated synchronization rates.

Each level creates its own type of fluctuations on throughput. The flow level fluctuations have already been studied both for throughput aggregates in [4, 13] and for the throughput of individual flows in [8]. For studying the fluctuations at network level, we introduced a representation of TCP controlled networks as billiards. We produced numerical evidence that both periodic and non-periodic asymptotic behaviors are possible for the empirical mean values of throughputs and that any slight

changes in the model parameters, for instance trends in the population parameters $a_{s,r}$ or n_s, could result into drastic changes for the instantaneous values achieved by empirical averages. Although flow level variability seems to be enough to provide a multifractal short time scale regularity, the combination of this and the network level fluctuations seems desirable in order to predict the global statistical structure of aggregated traffic.

This approach, together with its non-linear extension, was also shown to provide an efficient framework allowing one to simulate sizable networks. We intend to continue exploring this class of models and to try to enrich it with other types of traffic than the long lived TCP sessions on which this first step is focused.

References

[1] Abry, P. and Veitch, D. http://www.emulab.ee.mu.oz.au/∼darryl

[2] Abry, P., Flandrin, P., Taqqu, M.S. and Veitch, D. (2000) Wavelet for the analysis, estimation and synthesis on scaling data. *Self Similar Traffic Analysis and Performance Evaluation*, Park, K. and Willinger, W. Eds, Wiley.

[3] Baccelli,F., Bonald, T. (1999) Window flow control in FIFO networks with cross traffic. *Queueing Systems*, 32, 195-231.

[4] Baccelli, F. and Hong, D. (2002) AIMD, Fairness and Fractal Scaling of TCP Traffic. *Proc. of INFOCOM*, New York, July 2002.

[5] Bohacek, S. Hespanha, J. P., Lee, J. and Obraczka K. (2001) A Hybrid Systems Framework for TCP Congestion Control, Technical Report, USC, Los Angeles, CA, July.

[6] Bonald, T. and Massoulié, L. (2001) Impact of fairness on Internet performance. *SIGMETRICS*, pp. 82-91.

[7] Bu, T. and Towsley, D. (2000) Fixed Point Approximation for TCP Behavior in an AQM Network. *Proc. ACM SIGMETRICS*, vol. 29, no. 1, pp. 216-225.

[8] Chaintreau, A. and De Vleeschauwer, D. (2002) A Closed Form Formula for TCP Traffic Performance, *INRIA report*, to appear.

[9] Feldmann, A., Gilbert, A.C., Huang, P. and Willinger, W. (1999) Dynamics of IP traffic: A study of the role of variability and the impact of control. *Proc. of ACM-SIGCOMM'99*, August-September, Cambridge, MA, pp. 301-313.

[10] Gibbens, R. and Kelly, F. (1999), Resource Pricing and the Evolution of Congestion Control. *Automatica*, 35, pp. 1969-1985, 1999.

[11] Graham, C., Méléart, S. (1994) Chaos Hypothesis for a System Interacting through Shared Resources, *Probability Theory and Related Fields*, 100(2), pp.157-174.

[12] Guo, M., Crovella, M. and Matta, I. (2000) TCP Congestion Control and Heavy Tails. *Technical Report*, BUCS-TR-2000-017, Boston University.

[13] Hong, D. and Lebedev, D. (2001) Many TCP User Asymptotic Analysis of the AIMD Model. *Technical Report*, RR-4229, INRIA Rocquencourt.

[14] Hurley,P., Le Boudec, J.Y., Thiran, P. (1999) A Note on the Fairness of Additive Increase and Multiplicative Decrease. *Proc. of ITC-16*, Edinburgh, June.

[15] Kelly, F., Maulloo, A. and Tan, D. (1998) Rate control for communication networks: shadow price, proportional fairness and stability. Journal of the Operational Research Society, 49, pp. 237-252.

[16] Lakshman, T.V., Madhow, U. (1997) The performance of TCP/IP for networks with high bandwidth-delay products and random loss. *IEEE/ACM Trans. on Networking*, 5-3, pp. 336-350.

[17] Low, S.H. (2000) A Duality Model of TCP and Queue Management Algorithms. *Proc. of ITC Specialist Seminar*, CA, September.

[18] Low, S.H., Paganini, F., Wang J., Adlakha S., Doyle J. (2001) Dynamics of TCP/AQM and a Scalable Control, Proceedings of the 2001 Allerton Conference, University of Illinois, Oct. 2001.

[19] Massoulié, L. and Roberts, J. (1999) Bandwidth sharing: objectives and algorithms. *Proc. of INFOCOM*, New York.

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

[21] Mo, J. and Walrand, J. (2000) Fair End-to-End Window-based Congestion Control. *IEEE/ACM Trans. on Networking* 8-5, pp. 556-567.

[22] Mori, M. (1995) Zeta Functions and Perron-Frobenius Operator for Piecewise Linear Transformations on $\mathbb{R}^k$. *Tokyo Journal of Mathematics* 18, pp.401-416.

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

[24] Riedy R. and Levy-Vehel, J. (1996) Multifractal Properties of TCP Traffic. *Technical Report*, RR-3129, INRIA Rocquencourt.

[25] Roberts, J. and L. Massoulié. (1998) Bandwidth sharing and admission control for elastic traffic. *ITC Specialist Seminar*, Yokohama, October.

[26] Sinai, Ya. G. (1994). Topics in Ergodic Theory, *Princeton Mathematical Series*.

[27] Veres, A. and Boda, M. (2000) The Chaotic Nature of TCP Congestion Control, *Proc. of IEEE INFOCOM*, Tel Aviv.

[28] Vojnovic, M., Le Boudec, J.Y., Boutremans, C. (2000) Global fairness of additive-increase and multiplicative-decrease with heterogeneous round-trip times. *Proc. of IEEE INFOCOM*, Tel Aviv.

[29] Willinger, W., Paxson, V. and Taqqu, M.S. (1998) Self-Similarity and Heavy Tails: Structural Modeling of Network Traffic. *A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions*, R. Adler, R. Feldman and M.S. Taqqu (Eds.), Birkhauser Verlag, Boston, MA, pp. 27-53.

[30] Willinger, W., Taqqu, M.S., Sherman, R. and Wilson, D.V. (1997) Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level. *IEEE ACM Transactions on Networking*, Vol.5, No.1, pp. 71-86.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifi que
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399