A Mean-Field Analysis of Short Lived Interacting TCP Flows [*]

François Baccelli[†] Augustin Chaintreau[†] Danny De Vleeschauwer[‡] David McDonald[††]

[†] INRIA & ENS
45, rue d'Ulm
75005 Paris FRANCE
francois.baccelli@ens.fr,
augustin.chaintreau@ens.fr

[‡] Alcatel NSG
Francis Wellesplein 1
B-2018 Antwerpen Belgium
danny.de_vleeschauwer@alcatel.be

[††] University of Ottawa
585 King Edward Avenue
Ottawa, K1N6N8 Ontario
dmdsg@uottawa.ca

ABSTRACT

In this paper, we consider a set of HTTP flows using TCP over a common drop-tail link to download files. After each download, a flow waits for a random think time before requesting the download of another file, whose size is also random. When a flow is active its throughput is increasing with time according to the additive increase rule, but if it suffers losses created when the total transmission rate of the flows exceeds the link rate, its transmission rate is decreased. The throughput obtained by a flow, and the consecutive time to download one file are then given as the consequence of the interaction of all the flows through their total transmission rate and the link's behavior.

We study the mean-field model obtained by letting the number of flows go to infinity. This mean-field limit may have two stable regimes : one without congestion in the link, in which the density of transmission rate can be explicitly described, the other one with periodic congestion epochs, where the inter-congestion time can be characterized as the solution of a fixed point equation, that we compute numerically, leading to a density of transmission rate given by as the solution of a Fredholm equation. It is shown that for certain values of the parameters (more precisely when the link capacity per user is not significantly larger than the load per user), each of these two stable regimes can be reached depending on the initial condition. This phenomenon can be seen as an analogue of turbulence in fluid dynamics: for some initial conditions, the transfers progress in a fluid and interaction-less way; for others, the connections interact and slow down because of the resulting fluctuations, which in turn perpetuates interaction forever, in spite of the fact that the load per user is less than the capacity per user. We prove that this phenomenon is present in the Tahoe case and both the numerical method that we develop and simulations suggest that it is present in the Reno case too. It translates into a bi-stability phenomenon for the finite population model within this range of parameters.

Categories and Subject Descriptors

H.3.4 [**Distributed systems, Performance evaluation**]

General Terms

Performance, Theory

1. INTRODUCTION

Modeling TCP through the fairness it achieves (or equivalently the utility functions that it optimizes) has been a very active area of research since the work of Kelly in [15]. A general extension of this framework to dynamic traffic with a large number of flows is described in [10]. In [18] this framework is used to study the performance of networks with dynamic traffic (in [18] files to be transmitted arrive according to a Poisson process), with several types of fairness assumptions. In [12] the results proven in this previous paper are extended to a Poisson arrival process of sessions, each associated with a file download having a general distribution. [12] contains a proof that if the network can be modeled by a processor sharing (PS) queue - or equivalently if instantaneous fair sharing can be assumed in the network - then the mean throughput only depends on the average requested size per session. Comparison with simulations is provided but as the authors themselves remarked, this result might be challenged in real networks either for very small flows, that do not last long enough to benefit from their possible fair share in the network capacity, or for close to critical load where the discrimination between flows and the unequal sharing due to TCP are more frequent.

At the same time, a few papers focused on TCP bandwidth sharing for dynamic traffic when taking into account the AIMD rule. In [14], one of the first models developed on

[*]This research was supported in part by the "Opération Stratégique Conjointe" Alcatel-INRIA entitled "End to End Analysis of IP Traffic".

SIGMETRICS/Performance'04, June 12–16, 2004, New York, NY, USA.

dynamic traffic, a version of the Engset model is proposed and shown to be insensitive w.r.t the file size distribution. TCP is modeled as a constant transfer rate calculated from the study of TCP sharing for a fixed number of persistent flows that are exactly in phase (increasing their window and decreasing it by the same amount at the same time). This model is extended by Kherani and Kumar under exponential assumptions in [16] where the inter-congestion period and the increase of the total rate is now dynamically changing with the traffic. In this model the flows contributing to the traffic are all in phase (they all react together at the same time and in the same way); the analytical result cannot be explicitly given in the general case but only in the low load, large file case where TCP bandwidth can be approximated by a completely fair allocation.

Our work extends these two papers and proposes a new simple model for interacting HTTP flows sharing a common link. We are not assuming that the flows are in phase or that they share the bandwidth equally. We study the asymptotics of a model with N ON/OFF flows sharing a link according to TCP, when N tends to infinity. In the ON/OFF source model each source alternates between OFF periods and file transfers where both the file sizes and the OFF periods are independent random variables, with given distribution functions on the positive real line. We will consider two cases: the Reno case based on the additive increase multiplicative decrease (AIMD) rule for the transmission rate and the Tahoe case (the Reno case will be the default assumption throughout the paper).

It is well known that within the context of the Internet, it is appropriate to assume that the distribution of file sizes and OFF periods have heavy tails (e.g. Pareto file sizes and Weibull or lognormal OFF periods, as for example in [10]). However, in the mathematical part of the present paper, we will assume heavy tails because we are unable to solve the associated mathematics at this stage (nor can the rest of the scientific community to the best of our knowledge). We will rather concentrate on the version of the problem where both file sizes and OFF periods are exponential random variables and where all files and OFF times are mutually independent.

Why study a model based on statistical assumptions that are clearly inappropriate? The rationale is as follows: the exponential case is tractable and allows one to identify and prove the presence of phenomena that are also observed by simulation in the heavy tailed case. So the mathematical study based on the exponential case will be important step in the direction of the understanding of the interaction of HTTP flows with the more realistic statistics.

In the mathematical analysis, we assume the existence of a stationary deterministic mean-field limit when the number of flows goes to infinity. In this deterministic limit there are two possible stable regimes. If the file sizes are small enough the link is able to carry all the traffic without congestion. The average transmission rate stabilizes at a value calculated below giving an overall utilization of the link which is less than one. In the other stable regime there is a series of congestion epochs where the buffer overflows and the active flows experience losses and cut their transmission rate in two. The main aim of this paper is to investigate these two regimes, and in particular the conditions under which they appear and the stationary distributions they lead to.

Section 2 gives a necessary condition for the existence of stationary regimes with congestion epochs. This necessary condition is based on the rate conservation principle which allows one to pose a fixed point problem for the rate of congestion epochs. The numerical aspects associated with this fixed point equation are discussed in detail in this section: the functions that are used in this fixed point equation are obtained as the solutions of Fredholm integral equations of the second kind, which are derived from a regenerative analysis of the rate of a tagged flow. This leads to an efficient way of calculating the possible values of the period of the mean-field model.

Section 3 focuses on a necessary and sufficient condition for the existence of stationary regimes with congestion epochs. For this, we first study the interaction-less regime, for which we establish a partial differential equation. We give both an explicit solution of this PDE and an efficient numerical way to solve it via yet another Fredholm equation of the second type, which has a natural regenerative interpretation. We then establish an invariant equation describing, for a given inter-congestion period of the mean-field process, the stationary distribution of rates at a congestion epoch. The existence of a probability measure solution of this invariant measure equation which satisfies the certain conditions (explained in the paper) is a necessary and sufficient condition for the existence of such a periodic congestion regime. The associated integral equation is again a Fredholm integral equation of the second kind.

One of the key observations is made in Section 3 : within this setting, it is possible to have multiple stationary mean-field regimes depending on the initial conditions: for certain values of the parameters, there exist both a "fluid regime" where flows do not interact at all and a "turbulent regime" where the fact that flows interact once implies a slow down of the whole system that propagates interaction forever.

Section 4 extends the approach to a model with a simple representation of slow start. Section 6 gathers simulation results on the bi-stability phenomenon and on the case with heavy tailed file sizes and OFF-times. We show by simulation and analysis that the phenomena that are identified in the exponential model are also present in the heavy tailed case. Section 5 focuses on the comparison of our results with those of earlier models of the literature. In particular, we compare this model to the PS Engset model.

2. A NECESSARY CONDITION FOR THE EXISTENCE OF A REGIME WITH PERIODIC CONGESTION

2.1 Model

We suppose N HTTP flows share a link which has *no* buffer or rather a small buffer that cushions collisions. The link rate is CN packets per second so the link drops packets at random when the transmission rates of the flows exceed the link rate. We assume each HTTP flow is silent for an exponential time with a mean $1/\beta$. After the silence period the flow transmits a file where the distribution of file sizes is exponential with a mean $1/\mu$. The default option is that each flow implements TCP Reno so the transmission rate increases at rate $1/R^2$ during the transmission of a file where R is the round trip time of packets. When the file has been transmitted the transmission rate is reset to zero.

The interaction between flows is via the sum of their rates. As long as this sum, which we refer to as the aggregate rate, is less than NC, then there is no interaction between the flows. When the aggregate rate reaches the link capacity CN, an event that we call a congestion epoch occurs. For the sake of tractability, we assume that all losses taking place before the flows react take place instantaneously. This reaction consists in the fact that Reno may cut the rate given to each of the N flows independently with a probability p. The parameter p, which is the proportion of flows that experience a loss at such a congestion epoch, is called the synchronization rate of the model (this parameter is evaluated from queueing theory by Baccelli and Hong in [6]). After this reaction, the aggregate rate is again less than C and a new interaction-less phase starts. In the TCP-Tahoe case, the rate of flows that experienced a loss is reset to 0.

2.2 Rate Conservation

Define $X(t)$ to be the transmission rate of a tagged flow participating in the steady state. Assume that there exists a stationary regime for $X(t)$, namely that it is a stationary stochastic process defined on a probability space $\{\Omega, \mathcal{F}, P\}$. The distribution of $X(t)$ is therefore the distribution of all the transmission rates in the steady state. $X(t)$ increases linearly at rate $1/R^2$ when it is active; i.e. with mean rate $\mathbb{P}(X(0) > 0)/R^2$. This increase is counteracted by negative jumps when a file finishes and the transmission rate drops to zero. It is also counteracted by a reduction by one half when a packet is lost at a congestion epoch.

The following point processes will be useful:

- T, the point process of congestion epochs, with inter-arrival times τ, with Palm expectation $\mathbb{E}_0^\tau$; let $\bar{\tau}$ denote the expectation of the inter-congestion times w.r.t. $\mathbb{P}_0^\tau$;
- D, the point process of file completions of the tagged flow, with intensity λ_δ and with Palm expectation $\mathbb{E}_0^\delta$.

When a file is completely downloaded, the throughput is reset to zero. Hence, with the introduced notation, the rate of decrease of the transmission rate due to file completions is $\lambda_\delta \mathbb{E}_0^\delta(X(0^-))$. In addition to that, the mean rate at which the tagged flow suffers a packet loss is $p/\bar{\tau}$, and the tagged flow divides its transmission rate by 2 for each loss. Consequently the rate of decrease of the transmission rate due to packet loss is $\frac{p}{\bar{\tau}}\mathbb{E}_0^\tau[X(0-)/2]$. Since the utilization is exactly one when the congestion epoch begins it follows that $\mathbb{E}_0^\tau[X(0-)] = C$ so the rate of decrease of the transmission rate due to packet loss is $pC/(2\bar{\tau})$.

By the rate conservation principle (RCP, see e.g. [4], Chapter 1), the mean rate of increase equals the mean rate of decrease. So

$$\frac{\mathbb{P}(X(0) > 0)}{R^2} = \frac{pC}{2\bar{\tau}} + \lambda_\delta \mathbb{E}_0^\delta[X(0^-)]. \qquad (1)$$

On the left hand side the unknown quantity is the steady state probability that a flow is active while on the right hand side we have λ_δ, the rate at which file completions occur and $\mathbb{E}_0^\tau[X(0^-)]$, the mean transmission rate observed when the file is completely downloaded.

In the Tahoe case, the RCP equation reads

$$\frac{\mathbb{P}(X(0) > 0)}{R^2} = \frac{pC}{\bar{\tau}} + \lambda_\delta \mathbb{E}_0^\delta[X(0^-)]. \qquad (2)$$

In what follows, the RCP will be used as a way to determine the possible values of $\overline{\tau}$. As we shall see in §2.3 below the expressions that show up in the RCP equation, namely $\mathbb{P}(X(0) > 0)$ and $\mathbb{E}_0^\delta[X(0^-)]$ can be computed as a function of $\overline{\tau}$, so that this equation can be seen as a fixed point equation for $\overline{\tau}$.

2.3 The Fredholm Equations

In this section and in the rest of the paper, we let the parameter N tend to ∞ and we assume the existence of a stationary mean-field limit as $N \to \infty$ in the same spirit as in [7], [10] or [6]. In such a mean-field regime the inter-congestion times become deterministic and we have propagation of chaos; i.e. each flow becomes independent. We will concentrate on the case where the stationary regime of the mean-field limit has inter-congestion times are all equal to some constant τ.

We will see below that when assuming τ known, all quantities in Equation (1) can be computed as the solutions of certain Fredholm integral equations, and that (1) can be used as fixed point for determining τ.

In this section, we assume τ to be given. We define a cycle to start at a congestion epoch where the tagged flow is idle. The cycle ends at the first congestion epoch when the flow is idle again. We use the following notation :

- Σ is the point process of congestion epochs where the tagged flow is idle, with inter-arrival times σ and with Palm expectation $\mathbb{E}_0^\sigma$.

The rationale for defining such cycles is that the sequence of successive cycles associated with the tagged flow is i.i.d. or in other words that the beginning of cycles are regeneration times for the tagged flow.

2.3.1 *Expected number of files in a cycle*

Define $f(t)$ to be the expected number of files that will be transmitted by the end of the current cycle given that the tagged flow is inactive at the current time t (where $0 \le t < \tau$). Also define $g(z)$ to be the expected number of files that will be transmitted by the end of the current cycle given that the current transmission rate of the tagged flow is z packets per second and that the current time is immediately after a congestion epoch.

Our goal is to evaluate $f(0)$ but we find $f(t)$ for all $t \in [0, \tau[$. Since the silence period has an exponential distribution we can condition on the time when the flow has a new file to transmit. There are two possibilities. Either the file arrives before the next congestion epoch at some time r where $t \le r \le \tau$ or it does not. If it has not arrived, the current cycle ends and $f(t) = 0$.

If it does, for a time r where $t \le r \le \tau$, we condition on the size y of the arriving file. There are again two cases. Either the transmission of this file is completed before the next congestion or there is some remaining data to be transmitted after the next congestion epochs. We are in the first case if we can transmit y packets in $\tau - r$ time units given that the flow starts out with transmission rate zero. Since the transmission rate increases at rate $1/R^2$ it will take t' time units to transmit y packets if $y = (t'/2)(t'/R^2)$, ; i.e. if $t' = R\sqrt{2y}$. Consequently y packets can be transmitted before the next congestion epoch only if $y \leq (\tau - r)^2/(2R^2)$. In this case we add one to the number of files transmitted during the current cycle plus a renewal term. We can summarize this first case by

$$\int_t^\tau \beta e^{-\beta(r-t)} \left(\int_0^{\frac{(\tau-r)^2}{2R^2}} \mu e^{-\mu y} dy (1 + f(r + R\sqrt{2y}) \right) dr.$$

In the second case the y packets cannot be transmitted before the next congestion epoch. In this case, which occurs with probability $\exp(-\mu(\tau - r)^2/(2R^2))$, we do not add one to the number of files transmitted, but only the expected number of files transmitted after the next congestion epochs. It depends on the throughput seen after congestion : by the congestion epoch the transmission rate of the tagged flow is $(\tau - r)/R^2$. There is probability p that the tagged flow suffers a packet loss which reduces the transmission rate to $(\tau - r)/(2R^2)$.

We can summarize the expected number of files that will be transmitted by the end of the current cycle given we are in this second case as

$$\int_t^\tau \beta e^{-\beta(r-t)} e^{-\mu\frac{(\tau-r)^2}{2R^2}} \left(pg(\tfrac{\tau-r}{2R^2}) + (1-p)g(\tfrac{\tau-r}{R^2})\right) dr.$$

We conclude that $f(t)$ is given by :

$$\int_t^\tau \beta e^{-\beta(r-t)} \Bigg\{ \int_0^{\frac{(\tau-r)^2}{2R^2}} \mu e^{-\mu y} (1 + f(r + R\sqrt{2y}))dy$$
$$+ e^{-\mu\frac{(\tau-r)^2}{2R^2}} \left(pg(\frac{\tau-r}{2R^2}) + (1-p)g(\frac{\tau-r}{R^2})\right) \Bigg\} dr. \quad (3)$$

By similar arguments (see [5]), $g(z)$ can be written :

$$\int_0^{z\tau+\frac{\tau^2}{2R^2}} \mu e^{-\mu y}(1 + f(R\sqrt{R^2z^2 + 2y} - R^2 z))dy$$
$$+ e^{-\mu(z\tau+\frac{\tau^2}{2R^2})} \left(pg(\tfrac{z+\frac{\tau}{R^2}}{2}) + (1-p)g(z + \tfrac{\tau}{R^2})\right). \quad (4)$$

Equations (3) and (4) constitute an integral equation of the Fredholm type for the pair (f, g).

2.3.2 The three unknowns of the RCP equation

One can get similar Fredholm equation for determining the pairs of functions $(h, i), (j, k)$ and (l, m) where:

- $h(t)$ is the expected cumulative time that the flow is active in the remaining time of the current cycle given that the tagged flow is inactive at the current time t with $0 \leq t < \tau$.
- $i(z)$ is the expected cumulative time that the flow is active in the remaining time of the current cycle given that the current time is immediately after a congestion epoch, and that the tagged flow is active with a transmission rate of z.
- $j(t)$ is the expected residual time before the end of the current cycle given that the tagged flow is inactive at the current time t with $0 \leq t < \tau$.
- $k(z)$ is the expected residual time before the end of the current cycle given that the current time is immediately after a congestion epoch, and that the tagged flow is active with a current transmission rate of z.
- $l(t)$ is the expected cumulative throughput reductions due to file completions from now to the end of the cycle given that the tagged flow is inactive at the current time t with $0 \leq t < \tau$.
- $m(z)$ is the expected cumulative throughput reductions due to file completions from now to the end of the cycle given that the current time is immediately after a congestion epoch, and that the tagged flow is active with a rate of z.

The knowledge of

- $\mathbb{E}_0^\sigma[K_B] := f(0)$, the mean number of births during a cycle (which is also the mean number of file completions during a cycle;
- $\mathbb{E}_0^\sigma[\int_0^\sigma 1_{X(t)>0} dt] = h(0)$, the mean cumulative ON time over a cycle;
- $\mathbb{E}_0^\sigma[\sigma] = j(0)$, the mean duration of a cycle and
- $\mathbb{E}_0^\sigma[\int_0^\sigma X(t-)D(dt)] = l(0)$, the mean cumulative throughput reductions due to file completions over a cycle

in turn determines the 3 unknowns of (1) since:

$$\begin{aligned} \lambda_\delta &= \frac{\mathbb{E}_0^\sigma[K_B]}{\mathbb{E}_0^\sigma[\sigma]} = \frac{f(0)}{j(0)} \\ \mathbb{E}_0^\delta[X(0-)] &= \frac{\mathbb{E}_0^\sigma[\int_0^\sigma X(t-)D(dt)]}{\mathbb{E}_0^\sigma[K_B]} = \frac{l(0)}{f(0)} \\ \mathbb{P}(X(0) > 0) &= \frac{\mathbb{E}_0^\sigma[\int_0^\sigma 1_{X(t)>0} dt]}{\mathbb{E}_0^\sigma[\sigma]} = \frac{h(0)}{j(0)}. \end{aligned}$$

Notice that the product $\lambda_\delta \mathbb{E}_0^\delta[X(0-)]$ which is used in (1) is equal to $\frac{l(0)}{j(0)}$ so that the (f, g) pair is actually not required for solving this fixed point equation.

2.4 Numerical Evaluation of the Fixed Point

In this section we present the method that we developed to numerically study the fixed point equation satisfied by τ. The main result is a common linear equation describing the integral equations for the pairs $(f, g), (h, i), (j, k), (l, m)$.

Each of the pairs of functions $(f, g), (h, i), (j, k), (l, m)$ satisfies a Fredholm equation of the second type where all equations share some common terms. It is shown in [5] that the general form of these equations is as follows: we look for a functions A, defined on $[0; \tau]$ and a function B defined on $[0; +\infty[$ such that they verify Equation (5). In this equation, $\kappa = \mu/(R^2)$ and the functions U and V are given in the following table for all 4 cases:

$A(t)$	$B(r)$	$U(r)$	$V(r)$
$f(t)$	$g\left(\frac{r}{R^2}\right) - 1$	1	0
$h(t)$	$i\left(\frac{r}{R^2}\right)$	$a_\tau(r)$	$b_\tau(r)$
$j(t)$	$k\left(\frac{r}{R^2}\right)$	$a_\tau(r)$	$\frac{1}{\beta} + b_\tau(r)$
$l(t)$	$m\left(\frac{r}{R^2}\right) - \frac{r}{R^2}$	$\frac{a_\tau(r) + \frac{p}{2}c_\tau(r)}{R^2}$	$\frac{b_\tau(r) + \frac{p}{2}d_\tau(r)}{R^2}$

with the functions $a_\tau, b_\tau, c_\tau, d_\tau$ defined as:

$$a_\tau(r) = \int_r^\tau e^{-\kappa\frac{(s-r)^2}{2}} ds \; ; \; b_\tau(r) = \int_0^\tau e^{-\kappa\frac{s^2+2sr}{2}} ds$$
$$c_\tau(r) = (\tau - r)e^{-\kappa\frac{(\tau-r)^2}{2}} \; ; \; d_\tau(r) = (r + \tau)e^{-\kappa\frac{\tau^2+2\tau r}{2}}.$$

$$\begin{aligned} A(t) &= \int_t^\tau \beta e^{-\beta(r-t)} \left(U(r) + \int_r^\tau \kappa(s-r) e^{-\kappa\frac{(s-r)^2}{2}} A(s)ds + e^{-\kappa\frac{(\tau-r)^2}{2}} \left(pB(\frac{\tau-r}{2}) + (1-p)B(\tau-r) \right) \right) dr \ . \\ B(r) &= V(r) + \int_0^\tau \kappa(r+s) e^{-\kappa\frac{s^2+2sr}{2}} A(s)ds + e^{-\kappa\frac{\tau^2+2\tau r}{2}} \left(pB(\frac{\tau+r}{2}) + (1-p)B(\tau+r) \right) \ . \end{aligned} \quad (5)$$

Let $(\Gamma(t), \tilde{\Gamma}(r))$ be the solution (A, B) of Equation (5) for $(U, V) = (1, 0)$, let $(\Theta(t), \tilde{\Theta}(r))$ denote the solution for $(U, V) = (a_\tau, b_\tau)$, and let $(\Delta(t), \tilde{\Delta}(r))$ be the solution of this equation for $(U, V) = (c_\tau, d_\tau)$. According to the last table, we have :

$\Gamma(t) = f(t)$; $\Theta(t) = h(t)$ and, as Equation (5) is linear,

$$\frac{1}{\beta}\Gamma(t) + \Theta(t) = j(t) \text{ and } \frac{\Theta(t) - \frac{p}{2}\Delta(t)}{R^2} = l(t).$$

We numerically solve Equation (5) in the following way. First, we set $B(r) = 0$ for $x > K\tau$. This is motivated by the fact that for physical reasons $B(r)$ has to decrease as r increases, a fact that can be proved mathematically, but we omit the proof here. Second, we discretize the functions $A(t)$ and $B(r)$ uniformly with a density of M samples per interval of length τ. So, the function $A(t)$ is approximated by a vector of M samples and $B(r)$ by a vector of KM samples. We stack both vectors and hence obtain a vector of dimension $(K+1)M$. Approximating the integrals in (5) by weighted sums of the samples of the functions, Equation (5) reduces to a matrix equation. Solving this matrix equation involves the inversion of a $(K+1)M \times (K+1)M$ matrix. The numerical error introduced in this procedure can be controlled by the choice of the parameters K and M (see [5]).

2.5 Determination of τ

As shown above, τ satisfies the following equation :

$$\frac{pC}{2\tau} + \frac{l(0)}{j(0)} = \frac{1}{R^2}\frac{h(0)}{j(0)} \quad \text{i.e.} \quad C = \left(\frac{\Delta(0)}{\frac{1}{\beta}\Gamma(0) + \Theta(0)} \right) \frac{\tau}{R^2}. \quad (6)$$

This form is valid both for the Reno and the Tahoe cases, for appropriate definitions of Θ and Γ. In Figure 1, we have computed the right-hand side of the rightmost equation in (6), which does not depend on C, as a function of τ for a fixed setting of the parameters $1/\beta = 2s, 1/\mu = 2000$ Pkts, $R = 100ms$, $p = 0.8$ On this plot we can see that if the link capacity is large enough there is no value of τ making this function vanish (here for $C = 290$ Pkts/s.). In this case, the only possible stable regime is congestionless. For smaller values of the capacity, we observe either two fixed points (e.g. for C=270 Pkts/s.) or one (e.g. for C=250 Pkts/s.). In the case with two solutions, we have several candidates for a stable regime, with different periods. In the next section we will present a method helping to distinguish between solutions that may be the inter-congestion time of a stable regime and other solutions. From Figure 1 we can conclude more:

- for all C-values above 273.4 Pkts/s. (283.3 in the Tahoe case), there are no intersections;
- for $263 < C < 273.5$ Pkts/s. ($263 < C < 283.3$ in the Tahoe case), there are two intersections and
- for $C < 263$ Pkts/s., there is only one intersection.

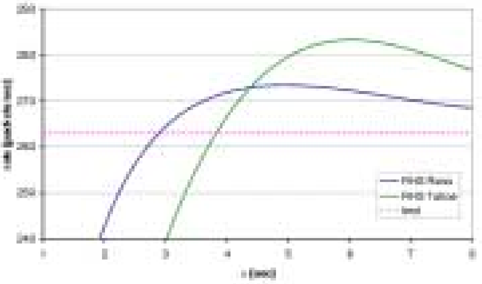

Figure 1: The RHS of (6) as a function of τ; the fixed points are the intersections of this RHS with the horizontal line C, in the Reno and the Tahoe cases.

2.6 The Tahoe Case

Given τ, the rate of the tagged flow is again a regenerative process with the same cycle structure as in the Reno case, namely starting with a congestion period when the rate of the tagged flow is 0 and ending at the next congestion is again 0. Using the same notation as in the Reno case, we now get

$$\begin{aligned} f(t) &= \int_t^\tau \beta e^{-\beta(r-t)} \\ &\cdot \left\{ \int_0^{\frac{(\tau-r)^2}{2R^2}} \mu e^{-\mu y}(1 + f(r + R\sqrt{2y})dy \right. \\ &\left. + e^{-\mu\frac{(\tau-r)^2}{2R^2}} \left(pg(0) + (1-p)g(\frac{(\tau-r)}{(R^2)}) \right) \right\} dr \end{aligned} \quad (7)$$

and

$$\begin{aligned} g(z) = &\int_0^{z\tau+\frac{\tau^2}{2R^2}} \mu e^{-\mu y}(1 + f(R\sqrt{R^2z^2 + 2y} - R^2z))dy \\ &+ e^{-\mu(z\tau+\frac{\tau^2}{2R^2})} \left(pg(0) + (1-p)g(z + \frac{\tau}{R^2}) \right). \end{aligned} \quad (8)$$

All other pairs can be analyzed in the same way (see [5]).

3. A NECESSARY AND SUFFICIENT CONDITION FOR THE EXISTENCE OF A CONGESTION REGIME WITH A GIVEN PERIOD

We start with a detailed study of the interaction-less regime (this is the free regime, i.e. the regime when $C = \infty$), which will be an essential ingredient of the analysis of the congestion regime which may occur when $C < \infty$ as we shall see in §3.2 below.

3.1 The Free Regime

3.1.1 The free regime regenerative rate process

In the case without congestion, each flow increases its transmission rate linearly at rate $1/R^2$ and can transmit a file of size y packets in time t where $y = t^2/(2R^2)$; i.e. in time $t = R\sqrt{2y}$. The density of the transmission time of a file is $\mu\frac{t}{R^2}e^{-\frac{\mu t^2}{2R^2}}$ (as easily seen by the change of variable $t \to v = t^2/2R^2$) and the mean file transmission time is therefore

$$T_{\mathtt{ON}} = R\int_0^\infty \mu\exp(-\mu y)\sqrt{2y}dy = R\sqrt{\frac{\pi}{2\mu}}. \tag{9}$$

A tagged flow alternates between periods composed of a silence period of exponential duration with parameter β and a active period of mean duration $T_{\mathtt{ON}}$, distributed according to the above density.

The rate $X(t)$ of the tagged flow at time t is a regenerative process that stays equal to 0 during OFF periods and increases linearly with time during activity periods. This stochastic process regenerates after the completion of one OFF and one ON period. The point process of regeneration epochs of a tagged flow will be denoted by S.

During each ON period a flow transmits on average $1/\mu$ packets. Consequently the average transmission rate per flow is

$$\rho = (1/\mu)/(1/\beta + T_{\mathtt{ON}})). \tag{10}$$

The proportion ν of flows which are idle is $(1/\beta)/(1/\beta + T_{\mathtt{ON}}))$. Notice that the transmission rate equals $\nu\beta/\mu$. This is intuitively obvious since $\nu\beta$ is the rate at which new flows come on-line and each new flow must transmit on average $1/\mu$ packets.

Hence when the regime without congestion occurs, the average transmission rate per flow ρ is less than C; i.e. $\nu\beta/\mu < C$ and

$$\rho = \frac{\nu\beta}{\mu} = \left(\mu\left(1/\beta + R\sqrt{\frac{\pi}{2\mu}}\right)\right)^{-1} < C. \tag{11}$$

3.1.2 The free regime PDE

Let $\nu(t)$ be the proportion of idle flows at time t. Let $s(z,t)$ be the density of the transmission rates of active flows in the mean-field regime (we consider first the case with a density for the sake of clear exposition). Consequently,

$$\int_0^\infty s(z,t)dz = 1 - \nu(t). \tag{12}$$

From the partial differential evolution equation introduced by Baccelli et al. in [7] we can see that the density function verifies the PDE:

$$\frac{\partial s}{\partial t}(z,t) + \frac{1}{R^2}\frac{\partial s}{\partial z}(z,t) = -\mu z s(z,t). \tag{13}$$

Multiplied by dz, the second term on the left hand side represents the rate of change of the proportion of transmission rates in $[z, z+dz]$ due to the linear increase in the transmission rate. The right hand side represents the rate at which files complete transmission since $s(z,t)dz$ is the proportion of flows with transmission rates in the interval $[z, z+dz]$ and flows with transmission rates in this interval complete transmission at a rate μz.

The rate at which flows become active is $\beta\nu(t)$ hence in time dt the area $\beta\nu(t)dt$ is added under the graph of $s(z,t)$ between 0 and dt/R^2 because this area is cleared out by the additive increase in the transmission rates. The area under the graph of $s(z,t)$ between 0 and dt/R^2 is $s(0,t)dt/R^2$ to first order. Hence,

$$s(0,t)/R^2 = \beta\nu(t). \tag{14}$$

It is shown in [5] using Laplace transform arguments that $s(z,t)$ satisfies the following Fredholm equation for $s(z,t)$:

$$\begin{aligned} s(z,t) \quad &= \quad s(z - \frac{t}{R^2}, 0)\ e^{-\mu\left(tz - \frac{t^2}{2R^2}\right)} + e^{-\mu R^2\frac{z^2}{2}}R^2\beta \\ &\left(1 - \int_0^\infty s(x, t - zR^2)dx\right) \end{aligned} \tag{15}$$

which turns out to be quite handy for numerical exploitation as we shall see below. Equation (15) is easy to interpret when considering the two cases: for the rate to be z at time t, either the transfer of the file transmitted at time 0 is not yet completed at time t, which requires that the rate was $z - t/R^2 \geq 0$ at time 0, or it is completed, which requires that the flow was inactive at time $t - zR^2 > 0$ and there was a transition from inactive to active at that time. In fact it is clear that (15) can be generalized to describe the evaluation of a measure $S(dz,t)$ representing the distribution of transmission rates at time t starting from an arbitrary measure $S(dz,0)$:

$$\begin{aligned} S(dz,t) \quad &= R^2\beta\left(1 - \int_{x=0}^\infty S(dx, t - zR^2)\right)\ e^{-\mu R^2\frac{z^2}{2}}dz \\ &+S(dz - \frac{t}{R^2}, 0)\ e^{-\mu(zt - \frac{t^2}{2R^2})}. \end{aligned} \tag{16}$$

Let

$$\alpha(t) = \int_0^\infty z s(z,t)dz. \tag{17}$$

The function $\alpha(t)$ represents the aggregate rate (sum of the transmission rates at time t where the sum is over all flows).

Lemma 1. The Laplace transform of $\alpha(t)$,

$$\widehat{\alpha}(u) \quad = \quad \int_0^\infty e^{-ut}\alpha(t)dt. \tag{18}$$

is

$$\widehat{\alpha}(u) \quad = \quad \frac{\nu(0)\frac{\beta}{\beta+u}\widehat{I}(u) + \widehat{J}(u)}{1 - \mu\frac{\beta}{\beta+u}\widehat{I}(u)}, \tag{19}$$

where

$$\widehat{I}(u) = R^2\int_0^\infty x e^{-R^2ux - R^2\mu x^2/2}dx \tag{20}$$

and $\widehat{J}(u)$ is given by

$$R^2\int_{z=0}^\infty e^{R^2uz + \frac{R^2\mu z^2}{2}}s(z,0)\int_{x=z}^\infty x e^{-R^2ux - \frac{R^2\mu x^2}{2}}dxdz. \tag{21}$$

LEMMA 2. The stationary distribution of the rates is:

$$\nu(\infty) = \frac{\frac{1}{\beta}}{\frac{1}{\beta} + R\sqrt{\frac{\pi}{2\mu}}}, \quad s(z, \infty) = \frac{R^2 e^{-R^2\mu z^2/2}}{\frac{1}{\beta} + R\sqrt{\frac{\pi}{2\mu}}}. \tag{22}$$

The stationary aggregate rate is:

$$\alpha(\infty) \quad = \quad \frac{1}{\mu}\frac{1}{\frac{1}{\beta} + R\sqrt{\frac{\pi}{2\mu}}} = \rho. \tag{23}$$

The proofs of these lemmas can be found in [5] where we also give an interpretation of the transforms of Lemma 1 in terms of renewal theory and a closed form expression for the solution of the PDE in the time domain.

3.2 The Interaction Regime(s)

3.2.1 The invariant measure equation

Assume there exists a periodic regime of period τ. Then τ should be a solution of (1). In addition the couple $(\nu_0, S_0(dz))$ that gives the proportion of OFF sources and the distribution of rates just after congestion epochs should be invariant w.r.t. the shift that moves from a congestion epoch to the next.

First τ and $(\nu_0, S_0(dz))$ should be such that the aggregate rate function α_0 obtained when taking $S(dz, 0) = S_0(dz)$ is such that $\alpha_0(\tau) = C$ and $\alpha_0(t) < C$ for all $0 < t < \tau$.

In addition, given that at congestion epochs, a proportion p of the windows are halved, the $(\nu_0, S_0(dz))$ should satisfy the integral equation (which will be referred to as the invariant measure equation)

$$S_0(dz) = (1-p)S(dz,\tau) + pS(d2z,\tau), \tag{24}$$

where $S(dz, t)$ is the solution of (16) with the initial condition $S(dz, 0)$ taken equal to $S_0(dz)$.

When using the explicit solution of the PDE given in the appendix of [5], one gets that the last integral equation for $S_0(.)$ can also be seen as a Fredholm type integral equation of the second kind.

In the Tahoe case the transmission rates of active sources has a measure which must have a point mass at zero at congestion epochs; the invariant measure equation then reads

$$S_0(dz) = (1-p)S(dz,\tau) + p\delta_0(dz)\int_0^\infty S(dv,\tau). \tag{25}$$

A few remarks are in order before addressing numerical issues:

- The existence of a couple $(\nu_0, S_0(dz)))$ solution of (24) and such that the $\alpha_0(\tau) = C$ and $\alpha_0(t) < C$ for all $t < \tau$ is necessary and sufficient for the existence of a congestion periodic regime of period τ. Using this, it is easy for instance to check that in the region where the RCP equation has two fixed points, the rightmost fixed point is spurious. This immediately follows from the fact that the condition $\alpha_0(t) < C$ for all $t < \tau$ is not satisfied for this other fixed point.
- The more general problem of finding all possible periodic regimes can be stated as follows: find all pairs made of a real number $0 < \tau < \infty$ and of a couple $(\nu_0, S_0(dz)))$ such that (24) (or (25) in the Tahoe case) holds and such that $\alpha_0(\tau) = C$ and $\alpha_0(t) < C$ for all $t < \tau$.
- Of course, other stationary regimes are possible like e.g. periodic regimes where the aggregate rate has a period that consists of $k > 1$ congestions, or even non periodic regimes (although we did not find such regimes by simulation).
- Injecting the couple $(\nu_0, S_0(dz)))$ as an initial condition into Equation (15) determines the proportion of active flows and the throughput distribution of active flows $S(dz, t)$ for all $0 \le t < \tau$. The mean stationary throughput obtained from this function averaged over continuous time is given by the following cycle mean:

$$\frac{1}{\tau}\int_{t=0}^{\tau}\int_{z=0}^{\infty} zS(dz,t)dt. \tag{26}$$

3.2.2 Numerical solution

We have chosen a numerical procedure to find an approximation for $s(z, t)$ based on Equations (15) and (24). We discretize the function $s(z, t)$ with $L + 1$ samples over its time domain (an interval of length τ) and with a density of L samples per interval of length $\frac{\tau}{R^2}$ over its space domain (i.e. the z variable). We use $L + 1$ samples in the time domain as there is a crucial difference between the time instant just before a congestion epoch (the L-th sample) and the time instant just after (the 0-th sample). We truncate the $s(z, t)$ function in the z direction by putting $s(z, t) = 0$ for $z > K\frac{\tau}{R^2}$. This truncation is motivated by the solution of the interaction-less system where this function decays like the tail of a Gaussian distribution.

The discretized version of Equations (15) and (24) define a matrix equation. Notice that in this case (in contrast to the case of solving for $A(t)$ and $B(r)$ in §2.4) there are L^2K unknowns and the matrices involved may become very large. Therefore, we used Equation (15) and (24) as a recursive rule to calculate an approximation for $s(z, t)$. The larger L and K are chosen the better the approximation (but more computations are needed). For the examples considered in this paper K=5 and L=200 turned out to be adequate values.

3.2.3 The multiple stationary regime region

In this section, we give both numerical and simulation evidence showing that the condition that the load factor

$$\rho = (1/\mu)/(1/\beta + T_{\mathtt{ON}}))$$

is less than C (namely the capacity per user is more than the mean load per user) is not sufficient for having an interaction-less mean-field regime for all initial conditions. The numerical part is based on the solution of the set of Fredholm equations of the last subsections. The simulation is based on the N2N code [2], a discrete event simulator which computes the AIMD sharing for a finite number of ON/OFF flows, interacting through the sum of their rates, as described in Section 2.1.

We also show that there exist values of the parameters such that depending on the initial condition describing the rates of the various flows, one may enter either into an interaction-less stationary regime or into a stationary congestion regime.

In the case considered here $1/\mu = 2000$ Pkts, $1/\beta = 2$ s., $p = 0.8$ and $R = 0.1$ s. The load factor ρ is then around 263 Pkts/s. We take $C = 270$ Pkts/s.

- When the initial condition is chosen according to the stationary law given in (22), then $\alpha(t) = \rho$ for all t and no congestions occur at all since $\rho < C$.

- As already shown in Section 2.5, the rate conservation principle gives two values of τ solution of the fixed point equation (1), the smallest of which is $\tau \sim 3.7s$. Using the solution of the invariant measure equation of Section 3.2.1, we find that for this value of τ, there exists a probability measure satisfying the integral equation (24) and satisfying the key condition that the associated α function first reaches C at time τ. The p.d.f of this distribution as obtained by two different methods is depicted in Figure 2 for Reno. The existence of such a regime is confirmed by the N2N simulation ([2]) of 1 Million HTTP users with the above characteristics and sharing a link of capacity 270 Pkts/s. Moreover, the steady state distributions found by simulation match quite precisely those obtained numerically.

In other words, depending on the initial phases of the flows, one either enters into a congestionless regime or into a periodic regime with infinitely many congestions. The first case occurs when the initial conditions are chosen independently for all flows, and each flow is in the stationary regime it would reach if there were no interaction at all. The second case occurs if the flows are more in phase: here all start inactive at time 0.

Here are a few remarks of interest:

- The same period and periodic regime are reached when the initial condition is that with all flows initially active and with null rate;

- The largest value of C for which we observe these two possible stationary regimes is approximately 273.5 Pkts/sec as shown independently by the N2N simulator and the fixed point method;

- the second solution of the RCP happens to be spurious. There exists a probability solution of (24) but the associated α function crosses the C level before this value of τ.

- Similar results hold for Tahoe. The associated distributions are given in [5].

3.2.4 Dependence of bi-stability region w.r.t. the parameters

Let C_T be the maximum C for which there is an interaction regime, ρ be given as in (10) and define the overprovisioning ratio (for guaranteeing the absence of interaction) to be $\omega = C_T/\rho$. Here are a few data on this ratio in the exponential case with $p = 0.8$ and $\frac{1}{\mu} = 2000$ Pkts.

- $1/\beta = 2$ s., $R = 0.1$ s.: $\omega = 1.04$;

- $1/\beta = 4$ s., $R = 0.1$ s.: $\omega = 1.06$;

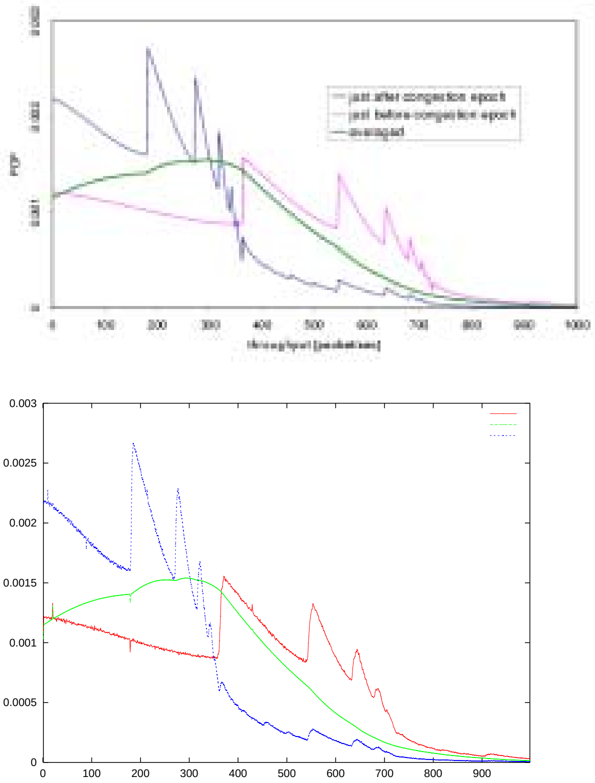

Figure 2: Distributions obtained for Reno. $1/\mu = 2000$ Pkts, $1/\beta = 2$ s., $p = 0.8$ and $R = 0.1$ s. and $C = 270$ Pkts/s. In red, steady state probability distribution function of the rate just after a congestion epoch; in green, continuous time stationary rate distribution. Top: numerical solution of the invariant measure equation; Bottom: N2N simulation of 1 Million HTTP flows.

- $1/\beta = 8$ s., $R = 0.1$ s.: $\omega = 1.09$;

- $1/\beta = 2$ s., $R = 0.05$ s.: $\omega = 1.06$;

- $1/\beta = 8$ s., $R = 0.05$ s.: $\omega = 1.12$;

- $1/\beta = 2$ s., $R = 0.025$ s.: $\omega = 1.09$;

- $1/\beta = 8$ s., $R = 0.025$ s.: $\omega = 1.15$.

The region is larger for small RTTs and for short think times.

3.2.5 Proof of the existence of congestion regimes with load less than capacity

Let us consider the Tahoe case with an initial condition consisting of all sources active and with 0 rate. The functions $\alpha(t)$ (the aggregate rate defined in (17)) and $\gamma(t) = 1 - \nu(t)$ (the proportion of active flows) associated with this initial condition play a key role in the construction of this section. They are depicted in Figure 3 in the case $1/\mu = 2000$, $1/\beta = 2$ and $R = 0.1$.

Let M denote the maximum of $\alpha(t)$ over all $t > 0$, θ the argmax of $\alpha(t)$, m the minimum of $\alpha(t)$ over all $t > \tau$ and let γ denote the minimum of $\gamma(t)$ over all $t > 0$. In the

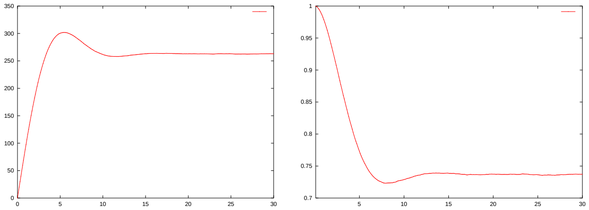

Figure 3: The α (left) and the γ (right) functions.

particular case of Figure 3, we have $M = 301.8$, $\theta = 5.5$, $m = 258.1$ and $\gamma = 0.723$. Let $\tilde{C} = p\gamma M + (1 - p\gamma)m$.

LEMMA 3. For the above initial condition, if $\tilde{C} > \rho$, then the Tahoe version of the model experiences an infinite number of congestion epochs for all C in the interval $\rho \leq C \leq \tilde{C}$.

For the proof, see [5]. So in our example, when $p = 0.8$, we are sure that Tahoe exhibits infinitely many congestions as soon as $C \leq \tilde{C} = 283.38$. Notice that this is only a sufficient condition for congestion, namely $\tilde{C} > C_T$ in general.

From our numerical and simulation estimates, it seems that the bi-stability region for Tahoe is larger than for Reno (see Figure 1).

Of course, under the assumption of the last lemma, if the initial condition for the flows is that of the steady state of the interaction-less regime, then one remains in this regime forever.

We have no analogue of Lemma 3 in the Reno case at this stage. The fact that Reno could have a turbulent regime when the load per user is less than the capacity per user is hence only backed by simulation and numerical evidence at this stage.

3.3 Properties of the Stationary Rate

We observe that a sharp decrease of the mean performance takes place at a value of the mean file size that is significantly smaller than that obtained by a mean load analysis. This sharp decrease is that due to the jump from the congestionless to the congestion stationary regimes described above (see the AIMD curve of Figure 5).

We now study more detailed properties of the stationary throughput. Figure 4 gives the stationary rate pdfs obtained by simulation and numerically in the case C=250 Pkts/sec, p=.4, $1/\mu$=2200 Pkts, $1/\beta$=2 s., R=.1 s. The fractal and intricate structure of the pdf of the rate at congestion epochs should not come as a surprise (similar shapes were obtained for long lived sessions by Chaintreau and De Vleeschauwer in [9]). Compared to the case of Figure 2 the irregularities of the pdf are enhanced by the smaller value of p. The continuous time stationary rate has a more regular rate pdf.

4. EXTENSION OF THE APPROACH TO THE SLOW START

4.1 Mathematical Analysis

The simplest way to represent slow start within this framework is via the an instantaneous jump of some random size at the birth of a flow. The rationale for this is that the associated exponential growth phase is quite quick compared to the congestion avoidance phase and that it can hence in a first approximation be seen as a jump from 0 to some random value H.

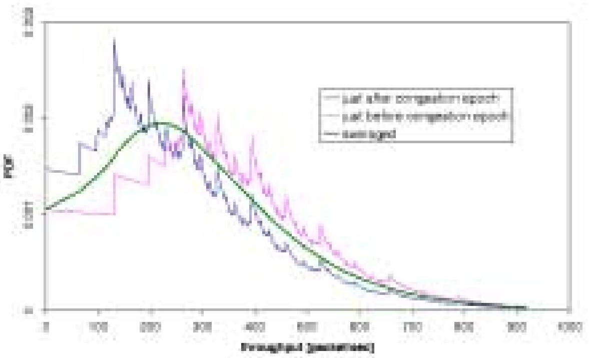

Figure 4: Reno case, C=250 Pkts/s., p=.4, $1/\mu$=2200 Pkts, $1/\beta$=2 s. Pdfs as obtained by the numerical method.

The RCP equation of Section 2.2 then becomes

$$\frac{pC}{2\tau} + \lambda_\delta \mathbb{E}^\delta(X(0-)) = \frac{P(X(0) > 0)}{R^2} + \lambda_\delta \mathbb{E}^B(H). \quad (27)$$

It is also easy to extend the integral equations of Section 2.3 from knowledge of the distribution $\eta(z)$ of H. The regenerative cycles admit the very same definition as in the case without slow start. The integral equations giving the expressions of $(f(.), g(.))$ and the other pairs can be found in [5]. So the fixed point equation based on the RCP can be extended almost directly to the case with this simplified representation of slow start. By arguments similar to the ones of §3.1 one gets (see [5]) that the solution of the associated PDE satisfies the Fredholm integral equation

$$\begin{aligned} s(z,t) &= R^2\beta \int_{v=0}^{z} \left(1 - \int_{x=0}^{\infty} s(x, t - R^2(z-v))dx\right) \\ &\qquad e^{-\mu R^2\left(\frac{z^2}{2} - \frac{v^2}{2}\right)} \eta(v) dv \\ &+ \quad s(z - \frac{t}{R^2}, 0)\ e^{-\mu\left(tz - \frac{t^2}{2R^2}\right)}. \end{aligned} \quad (28)$$

The invariant measure equation keeps the same form as (24) but with $s_0(.,.)$ now obtained from the last equation rather than from (15). The same machinery can then be used, in particular for the necessary and sufficient condition for the existence of a periodic regime of period τ, which is the direct analogue of what was done above in the case without slow start.

The numerical methods used for solving the RCP and the invariant measure equation and the simulation methodology are direct extensions of those used in the case without slow start.

Consider the case where H is deterministic and equal to $C/2$ (see §4.2 below).

Consider, for instance, the case where the parameters are still $C = 270$ Pkts, $1/\beta = 2$ s., p=0.8, R=0.1 s. and $1/\mu$=2000 Pkts. Both the N2N simulator and the RCP equation (27) give and a period of $\tau = 1.89$ s. The numerical solution of (28) leads to an aggregate rate function $\alpha(.)$ that satisfies the required property of first hitting $C = 270$ Pkts at $\tau = 1.89$ sec, so that this solution of the RCP is non-spurious.

4.2 HTTP 1.1 Example

We propose to focus on HTTP 1.1 where the files successively downloaded by a flow use the same TCP connection. This assumes of course that the successive downloads of this user are made from the same server and that the Keepalive Timer (usually 15 s.) does not expire (for the last point, see [1]).

We then refer to IETF RFC 2581 [3] to state the following concerning TCP:

- When the TCP connection is idle for more than one retransmission timeout (RTO, roughly a few RTTs), CWND is reduced to IW (initial window), which we will assume to correspond to decreasing the rate to 0.
- SSTHRESH is however kept to save information on the previous value of the congestion window. We propose here to take SSTHRESH= $C/(2(1-\nu))$, where ν denotes the stationary probability that a flow is idle at a congestion epoch. The rationale for this is as follows: when the last loss occurred (a loss always occurs for each flow in the finite population model), the proportion of active flows was $1-\nu$ and the average rate was per flow was C; hence due to symmetry, each active flow had an average of $C/(1-\nu)$; so it indeed makes sense to take SSTHRESH= $C/(2(1-\nu))$.

Hence in our slow start model, the rate of a flow jumps to $C/(2(1-\nu))$ at the beginning of each file transfer, and a congestion avoidance phase then starts until file completion. This is one model among many other possibilities, which has engineering meaning under the above assumptions (all flows access the same server, HTTP 1.1 is used, and the Keepalive Timer is large) and provided CWNDMAX is large and the exponential phase of the slow start is fast enough to be neglected.

Of course ν is unknown. To cope with this, in a first step we solve the model of §4.1 with $H = C/2$. This determines τ_1 and ν_1. In a second step, we solve the model again with $H = C/2(1-\nu_1)$ and so on until convergence. When applying this procedure to the example of the last section, τ_1=1.89 s. and ν_1=0.226 at the first step and τ_n=1.73 s. and ν_n=0.225 for all $n \geq 2$. The regime associated with the last values is such that the α function first reaches $C = 270$ Pkts at τ=1.73 s.

The basic observation is the same as in the case without slow start: in cases where the load per user is less than the capacity per user, one can get a turbulent mean-field limit with infinitely many congestions for appropriate initial conditions. Here is an example of such a turbulent regime: C=364 Pkts/s., p=.8, $1/\mu$=2000 Pkts, $1/\beta$=2 s. One gets a period of τ=5.568 s. and a load per user of 356.618 Pkts/s. Here, the load per user is defined using the same ideas as above: when the transfer of a file starts, the rate jumps from 0 to $H = C/(2(1-\nu))$ and then evolves according to the congestion avoidance AIMD rules. In this last expression, ν is the continuous time probability that a flow is active in the interaction-less regime. Notice that determining ν requires the solution of a fixed point equation (as this probability depends on H which itself depends on ν).

5. COMPARISON TO THE PS-ENGSET MODEL

An interesting issue concerning non persistent flows is the comparison of the bandwidth sharing that results from the AIMD induced dynamics of the present paper to that of the processor sharing (PS) approximations proposed in the literature (see Section 1). The closest large population PS model would be the Engset model with N users, where N is large. In this model the active sessions generate $1/\mu$ packets which are queued at a single server processor sharing node serviced with rate CN packets per second. Once served, these sessions move to an infinite server think time node where they stay for a duration of $1/\beta$ seconds. It is shown in [5] that when N tends to infinity, the mean rate obtained by each flow is $x = \beta \min\left(\frac{1}{\mu}, \frac{C}{\beta}\right)$. Figure 5 below compares this to the expressions obtained from our AIMD model, with and without slow start. In the case without slow start, the rate in the increasing part of the curve of the AIMD model (i.e. the part where no congestion occurs) is obtained from (10). As one can check, the match is not so good unless the load is small. Notice that there is actually no reason for these models to be close because, in the PS formula, there is no dependence on the RTT.

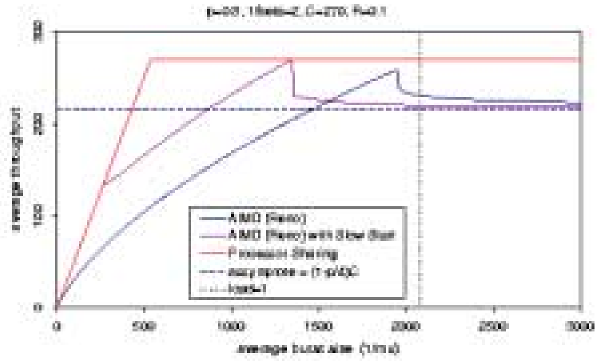

Figure 5: The average rate as predicted by the PS and the AIMD models.

The qualitative properties found in the present study have no analogues in these PS models:

- There are no multiple stationary regimes depending on the initial condition: above, we looked at the steady state of the Engset model and then let N (population) go to infinity. That is we let first time go to infinity (to get steady state) and we then let N to to infinity. Had we started the Engest model in some transient state (e.g. all users thinking, rather than in steady state), the steady state obtained when letting first N go to infinity and then letting time go to infinity is the same as the one obtained above as is easily seen by a direct analysis of the transient mean-field Engset model. Notice that these multiple regimes appear in the vicinity of critical load, which is precisely a region where PS is not expected to provide an accurate model for TCP bandwidth sharing anyway.

- The rightmost part of the PS curve postulates full bandwidth sharing whereas the AIMD dynamics does not. The rightmost part of the AIMD curve has an horizontal asymptote of app. $C(1-p/4)$ (that is here $.8 \times C$) as predicted by the long lived flow theory (see [6]). The abrupt drop in performance of about 15% when moving from the congestionless to the congestion regime is another qualitative feature (that is a consequence of the partial synchronization of flows) which is not present in the PS Engset model.

6. SIMULATION

The simulation results of this section are based on the N2N simulation tool [2].

6.1 Meta-stability

The fact that the mean-field limit has two stationary regimes for some values of the parameters translates into the existence of two meta-stable regimes for any finite stochastic system with the same mean parameters, with rare oscillations from one regime to the other. This phenomenon (see e.g. [13] for another example pertaining to protocols) is depicted in Figure 6 which features the Tahoe case with $1/\mu = 2000$ Pkts, $1/\beta = 2$ s. and $R = 0.1$ s.

In Figure 6, the number of sources is rather small (1000) and the capacity is approximately the critical value above which the mean-field system has only one uncongested mode. The two modes are clearly visible in the trajectories. The fluctuations are high enough to make the system move frequently enough from one mode to the other.

6.2 Heavy Tailed Case

The setting is the same as that of the previous sections with lognormal distribution functions for the file size and the OFF-time. The scenario is the following: TCP Reno, with RTT $R = 30$ ms. and with synchronization rate $p = 0.8$; the file size and the OFF-period follow lognormal distributions: the file size has mean value 2000 Pkts and standard deviation 8669 Pkts, and the OFF-period has a mean value of 2 sec and a standard deviation of 8.7 s. Variance is much higher than in the exponential case.

Simulations (or direct calculations) show that the mean load per source is appr. $\rho = 620$ Pkts/s. Figure 7 gives the aggregate rate when $C = \infty$ for the initial condition with all sources active and with null rate. We observe the same phenomenon as in the exponential case, with a first maximum at 717 Pkts/sec, significantly larger than the horizontal asymptote at ρ, though with a shape that is different from that in the exponential case.

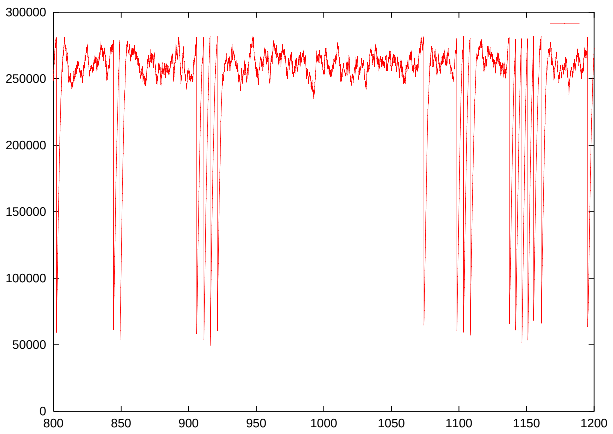

Figure 6: Bi-stability: 1000 Tahoe flows with $C = 282$, $p = .8$.

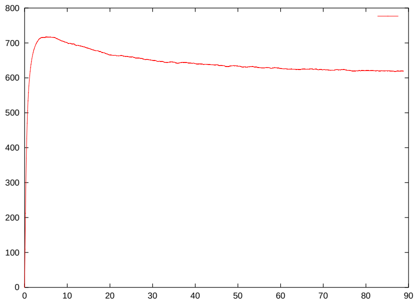

Figure 7: The mean-field aggregate rate of Reno when $C = \infty$ and all flows are initially active and with 0 rate.

Our simulation suggests that as in the exponential case, congestion regimes show up for values of C larger than ρ. Here, such regimes are possible for all C between a threshold that seems to be located between 670 and 680.

7. CONCLUSION

The main achievement of the present paper is an interaction model for TCP controlled dynamic flows that is based on the AIMD dynamics of TCP rather than on the frequently made assumption that TCP bandwidth sharing is well described by the PS discipline. Thanks to a mathematical model based on the mean-field limit, some unexpected qualitative results are found. In particular the system may enter into a congestion regime for loads that are significantly smaller than the link capacity. Also multiple stationary regimes may be reached depending on the initial phases of the ON-OFF flows. These phenomena, which translate into a bi-stability property for systems with finite population, are absent in the PS model.

Another interesting property is the fractal nature of the p.d.f of the stationary rates as already observed in the long-lived flow case by Chaintreau and De Vleeschauwer in [9]: the randomness and the mixing of the ON-OFF structure seems to be compatible with a complex self-similar structure for the rates. Even for the (rather unrealistic) exponential model analyzed here, several important theoretical questions have to be solved to complete the present study. These include the proof of the mean-field limit (which should be feasible along the lines of what was already done for the long lived flow case) and the mathematical confirmation of the numerical findings presented in Section 3 in the Reno case.

The main step after that is of course to extend the approach to non exponential file sizes and particularly to heavy tailed distributions. Other interesting extensions along the lines of what is already known for the long lived flow case would address the multiple link case and the nonlinear dynamics induced by a large tail-drop buffer. Finally, it should be possible to mix this HTTP traffic model with the model for long lived flows to give a single interactive, dynamical system.

Acknowledgments. The authors thank Dohy Hong for his valuable contributions to this paper. The analogy of the phenomena described in this paper with turbulence was brought to our attention by Bruce Hajek.

8. REFERENCES

[1] http://lists.w3.org/Archives/Public/ietf-http-wg-old/2000SepDec/0078.html

[2] http://www.n2nsoft.com

[3] http://www.faqs.org/rfcs/rfc2581.html

[4] F. Baccelli, P. Brémaud (2002), *Elements of Queueing Theory*, Springer Verlag, second edition.

[5] F. Baccelli, Chaintreau, A., De Vleeschauwer, D., McDonald, D. (2004) HTTP Turbulence, INRIA Report.

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

[7] Baccelli, F., McDonald, D. R., Reynier, J. (2002). A mean-field model for multiple TCP connections through a buffer implementing RED. *Performance Evaluation Vol. 11, (2002) pp. 77-97.* Elsevier Science.

[8] Barakat, C., Thiran, P., Iannaccone, C., Diot, C., Owezarski, P. (2002) A flow-based model for Internet backbone traffic. *Internet Measurement Workshop 2002.*

[9] Chaintreau, A., De Vleeschauwer, D. (2002) A closed form formula for long-lived TCP connections throughput. *Performance Evaluation 49(1/4): 57-76 (2000).*

[10] Chang, C.-S., Liu, Z. (2002) A Bandwidth Sharing Theory for a Large Number of HTTP-like Connections. *In Proceedings of the IEEE Infocom 2002 Conference, New York City , June 2002.*

[11] Feller, W. (1971) *An Introduction to Probability Theory and its Applications*, Wiley, second edition.

[12] Fred, S., Bonald, T., Proutiere, A., Régnié, G., Roberts, J. (2001) Statistical bandwidth sharing: a study of congestion at flow level. *in Proceedings of ACM SIGCOMM 2001: pp111-122*

[13] Gibbens, R. J., Hunt, P. J. and Kelly, F. P. (2002) Bistability in Communication Networks, *in Disorder in Physical Systems.*

[14] Heyman, D., Lakshman, T., Neidhardt, A. (1997) A new method for analyzing feedback-based protocols with applications to engineering web traffic over the Internet. *in ACM Sigmetrics, 1997, pp. 24–38.*

[15] Kelly, F. (1997) Charging and Rate Control for Elastic Traffic. *European Transactions on Telecommunications, vol. 8, pp. 33–37, 1997.*

[16] Kherani, A.A., Kumar, A. (2000) Performance Analysis of TCP with Nonpersistent Sessions. *Workshop on Modeling of Flow and Congestion Control, INRIA, Ecole Normale Supérieure, Paris, September 4-6, 2000.*

[17] Massoulié, L., Roberts, J. (1999) Bandwidth sharing: objectives and algorithms *in Proceedings of IEEE INFOCOM, Vol. 3. New York, NY, pp. 1395–1403.*

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