Investigating self-similarity and heavy tailed distributions on a large scale experimental facility

Patrick Loiseau, Paulo Gonçalves, Pascale Primet, Pierre Borgnat, Patrice Abry, Guillaume Dewaele

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Investigating self-similarity and heavy tailed distributions on a large scale experimental facility

Patrick Loiseau — Paulo Gonçalves — Pascale Primet Vicat-Blanc —
Pierre Borgnat — Patrice Abry — Guillaume Dewaele

N° 6472

March 2008

Thème NUM

ISSN 0249-6399 ISRN INRIA/RR--6472--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE – RHÔNE-ALPES

Investigating self-similarity and heavy tailed distributions on a large scale experimental facility

Patrick Loiseau*, Paulo Gonçalves* , Pascale Primet Vicat-Blanc* ,
Pierre Borgnat† , Patrice Abry† , Guillaume Dewaele‡

Thème NUM — Systèmes numériques
Équipe-Projet Reso

Rapport de recherche n° 6472 — March 2008 — 27 pages

Abstract: After seminal work by Taqqu et al. relating self-similarity to heavy tail distributions, a number of research articles verified that aggregated Internet traffic time series show self-similarity and that Internet attributes, like WEB file sizes and flow lengths, were heavy tailed. However, the validation of the theoretical prediction relating self-similarity and heavy tails remains unsatisfactorily addressed, being investigated either using numerical or network simulations, or from uncontrolled web traffic data. Notably, this prediction has never been conclusively verified on real networks using controlled and stationary scenarii, prescribing specific heavy-tail distributions, and estimating confidence intervals. In the present work, we use the potential and facilities offered by the large-scale, deeply reconfigurable and fully controllable experimental Grid5000 instrument, to investigate the prediction observability on real networks. To this end we organize a large number of controlled traffic circulation sessions on a nation-wide real network involving two hundred independent hosts. We use a FPGA-based measurement system, to collect the corresponding traffic at packet level. We then estimate both the self-similarity exponent of the aggregated time series and the heavy-tail index of flow size distributions, independently. Comparison of these two estimated parameters, enables us to discuss the practical applicability conditions of the theoretical prediction.

Key-words: Computer networks, Grid5000, metrology, self-similarity, heavy-tail distributions

* INRIA RESO, ENS Lyon, Université de Lyon.
† SiSyPhe, CNRS, ENS Lyon, Université de Lyon.
‡ SiSyPhe, ENS Lyon, Université de Lyon.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Vérification du lien entre auto-similarité et distributions à queues lourdes sur un dispositif grande échelle

Résumé : À la suite du travail théorique de Taqqu et de ses collaborateurs, reliant l'auto-similarité aux distributions à queues lourdes, quantité d'articles de recherche ont vérifié que les séries temporelles de trafic internet présentent en effet un caractère auto-similaire, et qu'en effet aussi, certaines variables d'internet, telle que par exemple les tailles de flux, étaient à queue lourde. Cependant, la validation de cette prédiction théorique liant auto-similarité et distributions à queues lourdes, reste peu satisfaisante dans la mesure où elle n'a été expérimentalement vérifiée que sur des simulateurs numériques de réseaux, ou sur des données de trafic réel dont on ne maîtrise aucun des paramètres. En particulier, cette relation n'a jamais été formellement validée sur des réseaux réels en situation contrôlée de scénarios stationnaires, dans lesquels des distributions à queues lourdes spécifiques sont prescrites, et des intervalles de confiances estimés. Dans ce travail, nous exploitons le potentiel et les capacités offertes par Grid5000, une plate-forme à grande échelle, profondément reconfigurable et totalement controlée, pour confronter cette prédiction théorique au contexte d'un véritable réseau. Pour ce faire, nous avons procédé à un grand nombre d'expériences *in situ*, où nous avons généré entre deux cents nœuds indépendants, différents profils d'un trafic entièrement contrôlé. Pour collecter les données correspondantes, nous utilisons un système à base de FPGA capable de traiter des flux de 1Gb/s avec une granularité à l'échelle du paquet. À partir de ces données, nous estimons indépendemment l'exposant d'auto-similarité du débit aggrégé et l'indice de queue lourde des distributions de taille de flux. La mise en correspondance de ces deux estimations nous permet alors de définir en pratique, les contours d'application du théorème.

Mots-clés : Réseaux informatiques, Grid5000, métrologie, auto-similarité, distributions à queues lourdes

1 Motivations

Comprehension and prediction of the network traffic is a constant and central preoccupation for internet service providers. Challenging questions, such as the optimization of network resource utilization that respect the application constrains, the detection (and ideally the anticipation) of anomalies and congestion, contribute to guarantee a better quality of service (QoS) to users. From a statistical viewpoint, this is a challenging and arduous problem that encompasses several components: network design, control mechanims, transport protocols and the nature of traffic itself. In the last decade, great attention has been devoted to the statistical study of time series and random variables, which collected at the core of networks, are valuable fingerprints of the system state and of its evolution. With this in mind, the pioneering work by [30] and [25] evidenced that the Poisson hypothesis, a relevant and broadly used model for phone networks, failed at describing computer network traffic. Instead, self-similarity was shown a much more appropriate paradigm, and since then, many authors have reported its existence in a wide variety of traffics [14, 23, 5, 6]. Following up this prominent discovery, the theoretical work by Taqqu and collaborators constituted another major breakthrough in computer network traffic modeling, identifying a plausible origin of self-similarity in traffic time series [25, 35, 37]. It is stated that the heavy-tail nature of some probability distributions, mainly that of flow size distributions, suffice to generate traffic exhibiting long range dependence, a particular manifestation of self-similarity [16]. To support their claim, they established a close form relation connecting the heavy tail thickness (as measured by a tail index) and the self-similarity exponent.

Notwithstanding its mathematical soundness, pragmatic validity of this model has been corroborated with real world traffic data only partially, so far. First pitfall lies in the definition of long range dependence itself, which, as we will see, is a scale invariance property that holds only asymptotically for long observation durations. Its consistent measurement requires that experimental conditions maintain constant, and that no external activity perturbs the traffic characteristics. In those conditions, finding a scale range that limits itself to stationary data, and that is sufficiently wide to endorse reliable self-similarity measurements, is an intricate task.

Secondly, even though real traffic traces had led to check concordance between tail index and self-similarity exponent, only was it perceived for a given network configuration that necessarily corresponded to a single particular value of the parameters set. An extensive test, to verify that self-similarity exponent obeys the same rule when the tail index is forced to range over some interval of interest, was never performed on a large scale real network plate-form.

Finally, the exact role of the exchange protocol, viewed as a subsidiary factor from this particular model, is still controversial [21, 28, 19]. Due to the lack of flexible, versatile, while realistic experimental environments, part of this metrology questioning has been addressed by researchers of the network community, using simulators, emulators or production platforms. However, these tools have limitations on their own, which turn difficult the studies, and yield only incomplete results.

In the present work, we use the potential and the facilities offered by the very large-scale, deeply reconfigurable and fully controlable experimental Grid5000 instrument to empirically investigate the scope of applicability of Theorem pro-

posed by Taqqu et al. [25, 35, 37]. Under controlled experimental conditions, we first prescribe the flow size distribution to different tail indices and compare the measured traffic self-similar exponents with their corresponding theoretical predictions. Then, we elucidate the role of the protocol and of the rate control mechanism on traffic scaling properties. In the course, we resort to efficient estimators of the heavy-tail index and of the self-similarity exponent derived from recent advances in wavelet based statistics and time series analysis.

The sequel is organized as follows. Section 2 summarizes related works. Section 3 elaborates on theoretical foundations of the present work, including a concise definition of parameters of interest. In section 4 we develop the specifities of our experimental testbed, and we describe our experimental designs. Section 5 presents and comments the results. Conclusions and perspectives are itemized in section 6.

2 Related Work

Without giving full bibliography on the subject (many can be found in [29, 5, 23]), there have been extensive reports on self-similarity in network traffic. As most of them are based on measurements and on analysis of real-world traces from the Internet, they only permit experimental validation of a single point on the curve, corresponding to one particular configuration. As its was presented before, the question here is more on the relation between these two properties, which is rooted in the seminal work by [25, 35] about the M/G/N queueing models with heavy-tail distributions of ON periods. Nonetheless, first experimental works by Crovella and co-authors [14, 28], hinted that this theoretical relation holds for internet traffic, and later on, also for more general types of traffic [21, 19]. However, due to the impossibility of controlling important parameters when monitoring the Internet, only compatibility of the formula could be tested against real data , but there is no statistically grounded evidences that self-similarity measured in network traffic is the work of this sole equality. On the other hand, study of self-similarity at large scales is very sensitive to inevitable non-stationnarities (day and week periodicities for instance) and to fortuitous anomalies existing on the Internet (see for instance [11]). It seems that the question has, since, never received a full experimental validation. In order to obtain such a validation, an important feature is to be able to make the heavy tail index vary, and there is only few attempts to validate the relation under these conditions. One is conducted in [28], that uses a network simulator, and where some departure from the theoretical prediction is reported (Fig. 3 in this article). This deviation is probably caused by the limited length of the simulation and also by the bias introduced by the used scaling estimator (R/S and Variance Time) on short traces. Actually, the main restriction of simulators lies in their scalability limitation, and in the difficulty of their validation. Indeed, the network is an abstraction, protocols are not production code, and the number of traffic sources or bitrates you can simulate depends on the computing power of the machine. Large-scale experimental facilities are alternatives that may overcome both Internet and simulators limitations as they permit to control network parameters and traffic generation, including statistics and stationarity issues.

Emulab [36] is a network experimental facility where network protocols and ser-

vices are run in a fully controlled and centralized environment. The emulation software runs on a cluster where nodes can be configured to emulate network links. In an Emulab experiment, the user specifies an arbitrary network topology, having a controllable, predictable, and reproducible environment. He has full root access on PC nodes, and he can run the operating system of his choice. However, the core network's equipments and links are emulated. The RON testbed [9] consists of about 40 machines scattered around the Internet. These nodes are used for measurement studies and evaluation of distributed systems. RON does not offer any reconfiguration capability at the network or at the nodes' level. The PlanetLab testbed [12] consists of about 800 PCs on 400 sites (every site runs 2 PCs) connected to the Internet (no specific or dedicated link). PlanetLab allows researchers to run experiments under real-world conditions, and at a very large scale. Research groups are able to request a PlanetLab slice (virtual machine) in which they can run their own experiment .

Grid5000, the experimental facility we use in the present work, proposes a different approach where the geographically distributed resources (large clusters connected by ultra high end optical networks) are running actual pieces of software in a real wide area environment. Grid5000 proposes a complimentary approach to PlanetLab, both in terms of resources and of experimental environment. Grid5000 allows reproducing experimental conditions, including network trafic and CPU usage. This feature warrants that evaluations and comparisons are conducted according to a strict and scientific method.

3 Theory

Taqqu's Theorem relates two statistical properties that are ubiquitously observed in computer networks: On the one hand, self-similarity that is defined at the level of aggregated time-series of the traffic, and on the other hand, heavy-tailness that involves grouping of packets (such as TCP connections). Simplistically, network traffic is described as a superposition of flows (without notions of users, or sessions,...) that permits us to adopt the following simple two-level model: *(i)* Packets are emitted and grouped in flows whose length (or number of packets) follows a heavy tailed distributed random variable [17, 10, 22]; *(ii)* the sum over those flows approximates network traffic on a link or a router. This crude description is coherent with current (yet more elaborate) statistical model for Internet traffic [10, 22].

After a succinct definition of these two statistical properties, we present the corresponding parameter estimation procedures that we use in our simulations, and chosen amongst those reckoned to present excellent estimation performance.

3.1 Self-similarity and long range dependence

3.1.1 Definition

Taqqu's Theorem implies that Internet time series are relevantly modeled by fractional Brownian motion (fBm), the most prominent member of a class of stochastic processes, referred to as *self-similar processes with stationary increments* (H-sssi, in short). Process X is said to be H-sssi if and only if its satisfies [16]:

$$X(t) - X(0) \stackrel{fdd}{=} X(u+t) - X(u), \forall t, u \in \mathbb{R}, \tag{1}$$

$$X(t) \stackrel{fdd}{=} a^H X\left(\frac{t}{a}\right), \forall t, a>0,\ 0<H<1, \tag{2}$$

where $\stackrel{fdd}{=}$ means equality for all finite dimensional distributions. Eq. (1) indicates that the increments of X form stationary processes (while X itself is not stationary). Essentially, self-similarity, Eq. (2), means that no characteristic scale of time can be identified as playing a specific role in the analysis or description of X. Corollarily, Eq. (2) implies that $\mathbb{E}X(t)^2 = \mathbb{E}X(1)^2 t^{2H}$, underlining both the scale free and the non stationary natures of the process.

It turns out that the covariance function of the increment process, $Y(t) = X(t+1) - X(t)$, of a H-sssi process X satisfies, for $|\tau| \to +\infty$:

$$\mathbb{E}Y(t)Y(t+\tau) \sim \mathbb{E}X(1)^2 H(2H-1)|\tau|^{2H-2}, \tag{3}$$

When $1/2 < H < 1$, hence $0 < 2-2H < 1$, such a power law decay of the covariance function of a stationary process is referred to as long range dependence [16, 29].

Long range dependence and self-similarity designate two different notions, albeit often confused. The latter is associated to non stationary time series, such as fBms, while the former is related to stationary time series, such as fBm's increments. In the present work, given that Taqqu's Theorem predicts that the cumulated sums of aggregated Internet time series are self-similar, we adopt here the same angle and discuss the results in terms of self-similarity of the integrated traces.

3.1.2 Self similarity parameter estimation

In [7], it was shown that wavelet transforms provide a relevant procedure for the estimation of the self-similarity parameter. This procedure revealed particularly efficient at analyzing Internet time series in [6, 5] and has then been massively used in this context.

Let $d_X(j,k) = \langle \psi_{j,k}, X \rangle$ denote the (Discrete) Wavelet Transform coefficients, where the collection $\{\psi_{j,k}(t) = 2^{-j/2}\psi_0(2^{-j}t-k), k \in \mathbb{Z},\ j \in \mathbb{Z}\}$ forms a basis of $L^2(\mathbb{R})$ [26]. The reference template ψ_0 is termed mother-wavelet and is characterized by its number of vanishing moments $N_\psi > 1$, an integer such that $\int t^k \psi_0(t)\, dt \equiv 0, \forall k = 0, ..., N_\psi - 1$. Then, decomposing a H-sssi process, the variance of the wavelet coefficients verifies [7]:

$$\mathbb{E}|d_X(j,k)|^2 = \mathbb{E}|d_X(0,0)|^2 2^{j(2H+1)}, \tag{4}$$

and, provided $N > H + 1/2$, the sequence $\{d_X(j,k),\ k = \ldots, -1, 0, 1, \ldots\}$ form a stationary and weakly correlated time series [6]. These two central properties warrant to use the empirical mean $S(j) = n_j^{-1} \sum_k |d_X(j,k)|^2$, ($n_j$ being the number of available coefficients at scale 2^j) to estimate the ensemble average $\mathbb{E}|d_X(j,k)|^2$. Eq. (4) indicates that self-similarity transposes to a linear behavior of $\log_2 S(j)$ vs. $\log_2 2^j = j$ plots, often referred to as Logscale Diagrams (LD) in the literature [6, 5]. A (weighted) linear regression of the LD within a proper range of octaves j_1, j_2 is used to estimate H.

In [6, 5, 7], the estimators performance are both theoretically and practically quantified, and are proved to compare satisfactorily against the best parametric techniques. Moreover, this estimator is endowed with a practical robustness that comes from its extra degree of freedom N_ψ. Its main use issue lies in the correct choice of the regression range $j_1 \leq j \leq j_2$. This will be discussed in Section 5, in the light of actual measurements.

3.2 Heavy Tail

3.2.1 Definition

A (positive) random variable $\mathbf{w}$ is said to be heavy tailed, with tail exponent $\alpha > 0$ (and noted α-HT) when the tail of its cumulative distribution function, $F_\mathbf{w}$, is characterized by an algebraic decrease [8]:

$$P(\mathbf{w} > w) = 1 - F_\mathbf{w}(w) \sim cw^{-\alpha} \quad \text{for } w \to \infty. \tag{5}$$

A α-HT random variable $\mathbf{w}$ has finite moments up to order α. For instance, when $1 < \alpha < 2$, $\mathbf{w}$ has finite mean but infinite variance. A paradigm for α-HT positive random variable is given by the Pareto distribution:

$$F_\mathbf{w}(w) = 1 - \left(\frac{k}{w+k}\right)^\alpha, \tag{6}$$

with $k > 0$ and $\alpha > 1$. Its mean reads: $\mathbb{E}\mathbf{w} = k/(\alpha - 1)$.

3.2.2 Tail exponent estimation

Estimation of the tail exponent α for α-HT random variables is an intricate issue that received considerable theoretical attention in the statistics literature: measuring the tail exponent of a HT-distribution amounts to evaluate from observations, how fast does the probability of rare events decrease in Eq. (5). Once random variables are know to be drawn from an a priori distribution, such as the Pareto form (6) for example, parametric estimators exist and yield accurate estimates of the tail index α (see e.g. [27]). However, if the actual distribution of observations does not match the a priori expected α-HT model, parametric estimators eloquently fail at measuring the tail decay.

For this reason, the non-parametric empirical estimator of α proposed in [20] will be preferred. The principle of this estimator is simple and relies on the Fourier mapping between the cumulative distribution function $F_\mathbf{w}(w)$ and the characteristic function $\chi_\mathbf{w}(s)$ of a random variable:

$$\chi_\mathbf{w}(s) = \int e^{-isw}\, dF_\mathbf{w}(w). \tag{7}$$

By a duality argument, the tail exponent α that bounds the order of finite moments of $F_\mathbf{w}$,

$$\alpha = \sup_r \{r > 0 : \int |w|^r\, dF_\mathbf{w}(w) < \infty\}, \tag{8}$$

transposes to the local Lipschitz regularity of the characteristic function $\chi_\mathbf{w}$ at the origin, according to:

$$\alpha = \sup_r \{r > 0 : 1 - \Re\chi_\mathbf{w}(s) = \mathcal{O}(s^r) \,\text{as}\, s \to 0^+\}, \tag{9}$$

where $\Re$ stands for the real part. It is easy to recognize in this power law behavior of $\Re\chi_{\mathbf{w}}(s)$, a scale invariance property of the same type of that of relation (3), which is conveniently identifiable with wavelet analyses. Hence, computing the discrete wavelet decomposition of $\Re\chi_{\mathbf{w}}$, and retaining only the wavelet coefficients that lie at the origin $k = 0$, yields the following multiresolution quantity:

$$d_{\chi_{\mathbf{w}}}(j,0) = \mathbb{E}\Psi_0\left(2^j\mathbf{w}\right) \leq C2^{j\alpha} \text{ for } j \to -\infty, \tag{10}$$

where $\Psi_0(\cdot)$ denotes the Fourier transform of analyzing wavelet $\psi_0(\cdot)$. Now, let $\{w_0, \cdots, w_{n-1}\}$ be a set of i.i.d. α-HT random variables, and replace the ensemble average in Eq. (10) by its empirical estimator, the estimate $\widehat{\alpha}$ simply results from a linear regression of the form

$$\begin{aligned} \log \widehat{d}^{(n)}_{\chi_{\mathbf{w}}}(j,0) &= \log n^{-1} \sum_{i=o}^{n-1} \Psi(2^j w_i) \\ &\approx \widehat{\alpha} j + \log C, \text{ as } j \to -\infty. \end{aligned} \tag{11}$$

The estimator was proven to converge for all heavy tail distributions, and also it has a reduced variance of estimation in $\mathcal{O}(n^{-1})$, where n is the sample size. We refer the interested reader to [20] where robustness and effective use of this estimator are thoroughly studied. Yet, let us mention the existence of a theoretical scale range where the linear model, Eq. (11), holds, and which shows very helpful for practitioners to adequately adjust their linear fitting over a correct scale range.

3.3 Taqqu's Theorem

A central result for interpreting statistical modeling of network traffic is a celebrated Theorem due to M. Taqqu and collaborators [25, 35, 37], in which heavy-tailness of flow sessions has been put forward as a possible explanation for the occurrence of self-similarity of Internet traffic.

The original result considers a M/G/N queueing model served by N independent sources, whose activities $Z_i(t)$, $i \in \{1, ..., N\}$, are described as a binary ON/OFF processes. The durations of the ON periods (corresponding to a packet train emission by a source) consists of i.i.d. positive random variables τ_{ON}, distributed according to a heavy-tail law P_{ON}, with exponent α. Intertwined with the ON periods, the OFF periods (a source does not emit traffic), have i.i.d. random durations τ_{OFF} drawn from another possibly heavy-tailed distribution P_{OFF} with tail index β. Thus, the $Z_i(t)$ consist of a 0/1 reward-renewal processes with i.i.d. activation periods.

Now, let $Y_N(t) = \sum_{i=1}^{N} Z_i(t)$ denote the aggregated traffic time series and define the cumulative process $X_N(Tt)$:

$$X_N(tT) = \int_0^{Tt} Y_N(u)\mathrm{d}u = \int_0^{Tt} \left(\sum_{i=1}^{N} Z_i(u)\right) \mathrm{d}u. \tag{12}$$

Taqqu's Theorem (cf. [35]) states that when taking the limits $N \to \infty$ (infinitely many users) and $T \to \infty$ (infinitely long observation duration), in this order, then $X_N(tT)$ behaves as:

$$X_N(tT) \sim \frac{\mathbb{E}\tau_{\text{ON}}}{\mathbb{E}\tau_{\text{ON}} + \mathbb{E}\tau_{\text{OFF}}} NTt + C\sqrt{N}T^H B_H(t). \tag{13}$$

In this relation, C is a constant and B_H denotes a fractional Brownian motion with Hurst parameter:

$$H = \frac{3 - \alpha^*}{2}, \text{ where } \alpha^* = \min(\alpha, \beta, 2). \tag{14}$$

The order of the limits is compelling to obtain this asymptotic behavior; this has been discussed theoretically elsewhere and is beyond the issues we address here. The main conclusion of Taqqu's Theorem is that, in the limit of (infinitely) long observations, fractional Brownian motions superimposed to a deterministic linear trend, are relevant asymptotic models to describe the cumulated sum of aggregated traffic time series. Moreover, Eq. (14) shows that only heavy tailed distributions with infinite variance (i.e., $1 < min(\alpha, \beta) < 2$) can generate self-similarity associated to long range dependence (i.e. $H > 1/2$). Conversely, when both activity and inactivity periods have finite variance durations, $\alpha^* = 2$ and consequently $H = 1/2$, which means no long range dependency.

4 Experimental setup

To study the practical validity of Taqqu's result, we use the potential and facilities offered by the very large-scale, deeply reconfigurable and fully controllable experimental Grid5000 instrument, so as to overcome the limitations previously exposed of emulations, simulations or measurements in production networks. After a general overview of Grid5000, the metrology platform is described first. Design of a large set of experiments, aimed at studying the actual dependence between the network traffic self-similarity and the heavy-tailness of flow size distributions, is finally detailed.

4.1 Grid5000 instrument overview

Grid5000, is a 5000 CPUs nation-wide Grid infrastructure for research in Grid computing [13], providing a scientific tool for computer scientists similar to the large-scale instruments used by physicists, astronomers, and biologists. It is a research tool featured with deep reconfiguration, control and monitoring capabilities designed for studying large scale distributed systems and for complementing theoretical models and simulators. Up to 17 french laboratories involved and 9 sites are hosting one or more cluster of about 500 cores each. A dedicated private optical networking infrastructure, provided by RENATER, the French NREN is interconnecting the Grid5000 sites. Two international interconnection are also available: one at 10 Gb/s interconnecting Grid5000 with DAS3 in Netherlands and one at 1 Gb/s with Naregi in Japan. In the Grid5000 platform, the network backbone is composed of private 10 Gb/s Ethernet links connected to a DWDM core with dedicated 10 Gb/s lambdas with bottlenecks at 1 Gb/s in Lyon and Bordeaux (see Figure 1).

Grid5000 offers to every user full control of the requested experimental resources. Its uses dedicated network links between sites, allows users reserving the same set of resources across successive experiments, allows users running their experiments in dedicated nodes (obtained by reservation) and lets users install and run their proper experimental condition injectors and measurements software. Grid5000 exposes two tools to implement these features : OAR is

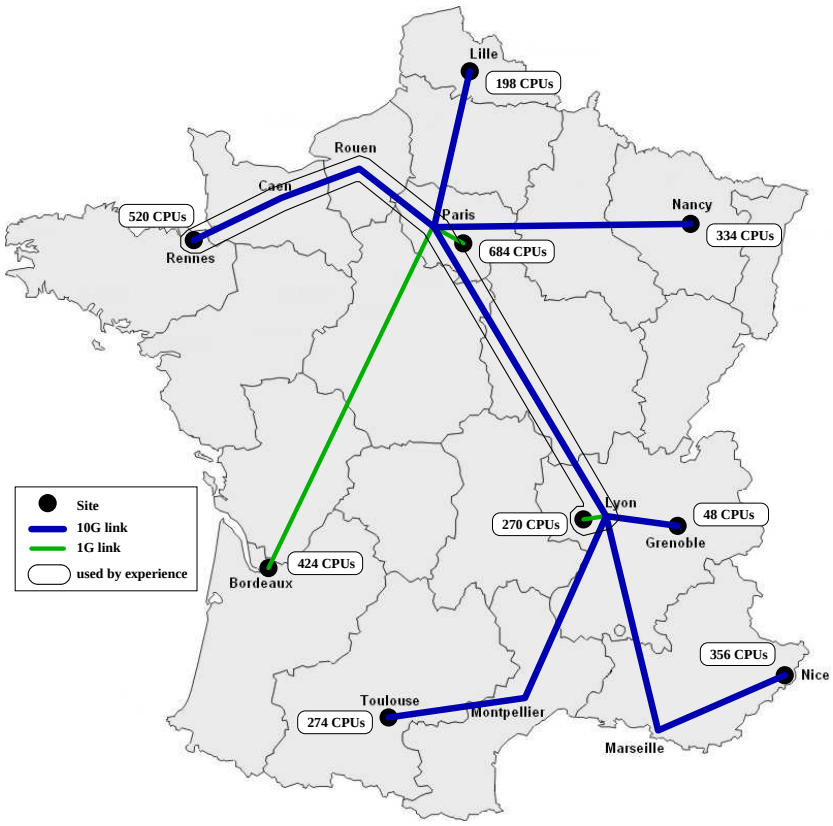

Figure 1: Grid'5000 backbone

a reservation tool, and Kadeploy an environment deployment system. OAR offers an accurate reservation capability (CPU/Core/Switch reservation) and integrates the kadeploy system. With Kadeploy, each user can make his own environment and have a total control on the reserved resources. For instance, software and kernel modules for rate limitation, QoS mechanisms, congestion control variants can be deployed automatically within the native operating system of a large number of communicating nodes. OAR also permits users to reserve equipements for several hours. As a consequence, Grid5000 enable researchers to run successive experiments reproducing the exact experimental conditions several times, an almost impossible task with shared and uncontrolled networks. This insures also large-duration observation windows under stationary conditions – something that is unachievable on the Internet. As a private testbed, Grid5000 turns the installation of experimental hardware, like for instance the traffic capture instrument at representative traffic aggregation points, quite easy.

4.2 Metrology platform

Using the facilities offered by Grid5000, a platform for metrology has been designed, an schematized in Fig. 2. Before describing the monitoring facilities

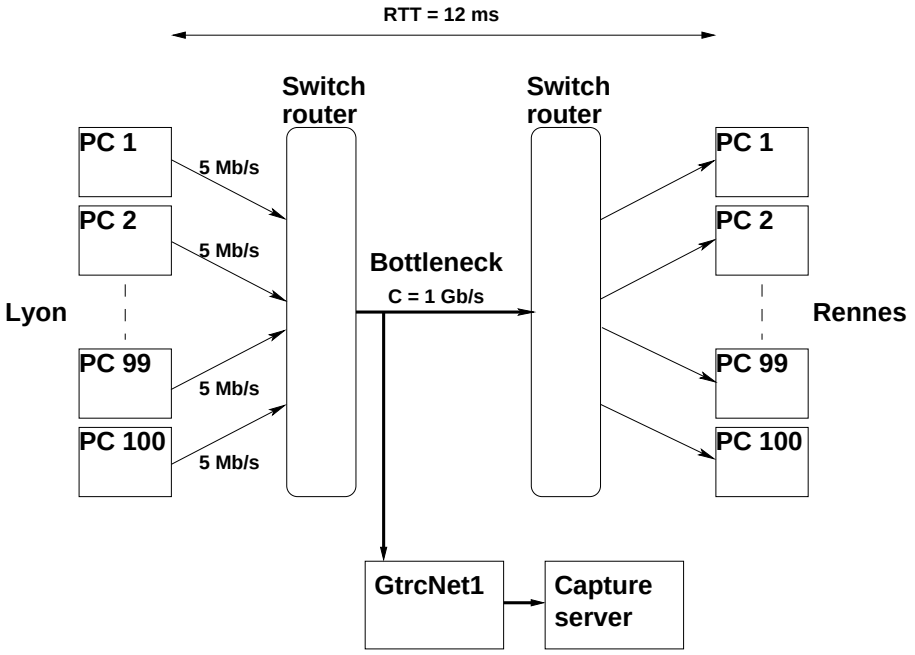

Figure 2: Metrology platform overview

and the developed data processing softwares, let us present the effective topology used for this experiment.

4.2.1 Experimental system description

Unless explicitly mentioned, all our experiments consist in producing data flow transfer between many independent client nodes (sources) and many independent server nodes (destinations). It is a classical dumbbell (or butterfly) topology with a single bottleneck of capacity, here of $C = 1$ Gb/s. We selected $N = 100$ nodes that are able to send up to at $C_a = 1$ Gb/s on each direction (see Fig. 2)

For the experiments, we used nodes on the Grid5000 clusters of Lyon (clients) and Rennes (server). The average RTT is then stable and equal to 12 ms which gives a bandwidth-delay product of 1.5 MBytes. In our forthcoming scenarios, TCP and UDP transfers are realized by using *iperf* [2] on Sun Fire V20z (biopteron) workstations of Grid5000 [13], running GNU/Linux 2.6.18.3 kernels with standard TCP and UDP modules. Iperf is a traffic generation tool that allows users to tune the different TCP and UDP parameters and to evaluate their impact on network performance.

4.2.2 Capture system

To measure the traffic at packet-level, we designed a specific system combining packet capture, header extraction and dedicated data analysing software. Packets are first captured by mirroring traffic of the access link connecting Lyon site

to the rest of Grid5000. Only the outgoing traffic from Lyon site to Grid5000 is mirrored, connecting a 1 Gb/s fiber port to a 1 Gb/s copper port directed to the monitoring system.

This system is composed of a GtrcNET-1 device [24], developed by AIST GTRC, and based on an FPGA that has been programmed to extract and aggregate packets headers and send them to an attached server. This header aggregation reduces the number of interrupts of the computer that receive the traffic to analyse, decreasing the local loss probability. In the packet capture system, the GtrcNET-1 is configured to extract a 52-Bytes headers (composed of 14, 20 and 18 Bytes of Ethernet, IP and TCP headers respectively) from the packets arriving at the one gigabit port. Headers are added a time-stamp each, encapsulated by groups of 25 into a UDP packet and then sent to another gigabit port. Time-stamps have a 60 ns ($2^{-24}s$) resolution.

The concatenated headers are stored in a computer with a quad core processor running at 2.66 GHz, 4 GB memory, 2 ethernet gigabit ports, 300 GB SAS disk for the system, 1 RAID controller with 5 x 300 GB SAS disk in a RAID 0 array offering 1500 GB available for storing capture files. We developed a driver that reads GtrcNET-1 packets, de-encapsulates time-stamped packet headers and writes them to a file in the pcap format.

4.2.3 Data processing and Flow Reconstruction

We use a series of tools over captured IP traffic traces, to go from the packet-level traces to the aggregated traffic and the flow statistics that are needed in this work. A first step is to handle the captured IP traffic traces; second, we reconstruct the flows from the packets[1].

IP traffic traces, saved in standard pcap format by the capture device, are first processed by `ipsumdump`, a program developped at UCLA [4], able to read the pcap format and to summarizes TCP/IP dump files into a self-describing binary format. Thanks to this tool, we retrieve from our traces the needed informations: timestamps, source and destination IPs, port numbers, protocol, payload size, TCP flags, and sequence numbers. The information are condensed into a binary file that is easier to parse, and which doesn't depend on specific capture hardware anymore.

Second, we have developed a collection of tools working on the `ipsumdump` binary format directly, which performs a variety of useful data operations on the traces. Of relevant interest here are: computation of the aggregated time-series of traffic (used for self-similarity estimation); extraction of traffic sub-traces for conditioned study, based on flows or packets random sampling, or on parameters filtering (traffic from/to a list of IPs, traffic on given ports, traffic using a specific protocol, etc); and reconstruction of the flows existing in the traces.

The question of flow reconstruction is an intricate problem, that is an important and difficult aspect when one wants to study the impact of their heavy-tailness [18, 15, 29, 31]. It is necessary to recompose each flow from the intertwined packets stream measured on an aggregated link. This means we must identify and group all the packets pertaining to the same set, while considering a significantly large number of flows to guarantee statistical soundness. This

[1] Using standard flow monitoring tools, such as Netflow or Sflow, would not be sufficient here. Indeed, we need statistical characterization at both the packet-level (for H-sssi) and the flow-level (for α-HT).

constraint faces the arduous issue of loss free capture, and that of dynamic table updating.

In our tool, flows are classically defined as a set of packets that share the same 5-uplet comprising: source and destination IPs, ports, and protocol. However, because there is a finite number of ports, it is possible for two different flows to share the same 5-uplet, and thus to get grouped in a single flow. To avoid this, we set a `timeout` threshold: a flow is considered as finished, if its packet train undergoes an interruption lasting more that `timeout`. Any subsequent packet with the same 5-uplet will tag the beginning of a new flow. Naturally, a proper choice of `timeout` is delicate, but that is the only solution that works for any kind of flows. For TCP flows, though, things are easier, as we can use the SYN or SYN/ACK flags to initiate a flow (closing any currently open flow with the same 5-uplet), and the FIN or RCT flag to close the flow, dispensing with `timeout`. Note that `timeout` remains necessary when the FIN packet is accidently missing.

Flow reconstruction (with `timeout`) is then performed in a table that contains all currently open flows, using hash functions to speed up the access. The relatively modest trace bitrate allows for keeping the whole table in memory. Since TCP sequence numbers and payload size for TCP packets are captured, it is possible to search for dropped or re-emmited packets during the flows' reconstruction and take that into account. Elementary statistics on the flows are then available: number of packets, number of Bytes, duration of the flow, etc.

All together the data processing tools extract the two elements needed for this study: the aggregated time-series at the packet-level, and the experimental flow-size distribution of any traffic that will be sent through, and monitored in the metrology platform of Grid5000.

4.3 Rate limitation mechanisms

The last major aspect of the experimentation is the careful design of traffic generation. In real network, flows are not the fluid ON/OFF flows of the $M/G/N$ model: packets composing the flows are sent entirely, one after the other, at the wire bit-rate. This acts as an ON/OFF sending process. Following on, a critical feature to consider in network experiment design, is the mechanism of traffic generation, especially the rate at which the packets are sent. An important parameter is then the aggregation level of the traffic K, defined as the ratio between the bottleneck capacity C and the access link nominal capacity C_a. In xDSL context and more generally in the Internet, it is not rare to have K ranging over 1000, while in the data-center context, K is around 1 or 10. In our Grid5000 setup the K factor is close to 1. To obtain a K factor larger than 100 and to insure an aggregated throughput average lower than $C = 1$ Gb/s, the sources rate has to be limited at most to 10 Mb/s.

End-host based mechanisms can control the individual flows in a scalable and simple way [33]. When considering fixed size packets, the way to modify data rates over a large period of time is to vary inter-packets intervals. To calculate these intervals, one considers the time source that can be used to enforce the limitation. In end-host systems, four different time sources are available: a) userland timers, b) TCP self clocking namely RTT of the transfer's path, c) OS's kernel timers, d) Packet-level clocking. These time sources allow to create different sending patterns as shown on figure 3. In our experiments we used

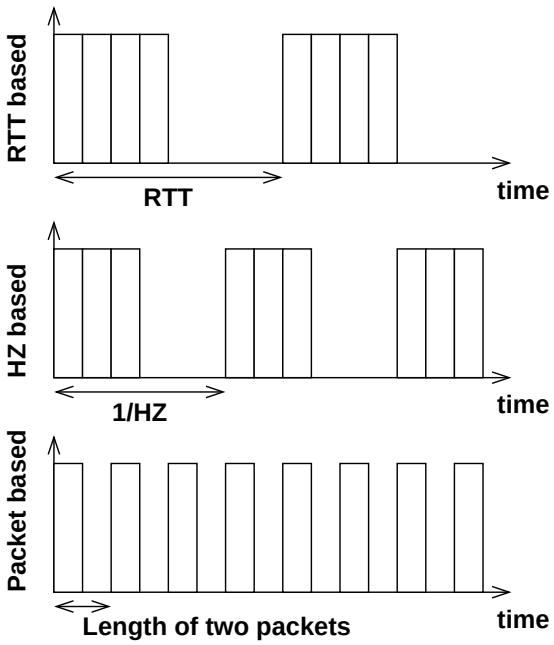

Figure 3: Different sending patterns. The upper figure shows a schematic view of an RTT-based rate limitation, the lower ones packet-level rate limitation and in-between timer based rate limitation.

three rate limitation approaches which act at different time scales: the first one is based on packet-level clocking (packet spacer), the second one on OS's kernel timers (Token Bucket), the last one on TCP self clocking namely RTT of the transfer's path (window size limitation).

The first two methods rely on the linux traffic shaping mechanism: with the `tc` utility [3], the qdisc associated to a network interface (the queue in which packets are put before being sent to the network card) is configured. The PSP (PSPacer) [34] qdisc spaces packets by inserting IEEE 802.3x PAUSE packets. These PAUSE packets are discarded at the input port of the first switches. With this mechanism, packets are regularly spaced and short bursts are avoided. The second method resorts to HTB (Hierarchical Token Bucket) qdisc [1] that uses a bucket constantly filled by tokens at the configured target rate. With this qdisc the average rate limit can be overridden during short bursts.

The third and last method modifies the TCP window size to slow down the throughput. The formula $window\_size = target\_throughput \times RTT$ determines the TCP window size to use to limit the sending rate to $target\_throughput$. This mechanism works well if the window size is not to small, which means also that the target throughput and/or the RTT should not be too small either. As the TCP limitation acts for each TCP connection, many sources located on the same node can have independent rate limitation which is not the case for qdisk-based limitation mechanisms. To limit the rate of a 1 Gb/s source to 5 Mbps with full size (1500 Bytes) Ethernet packet and RTT of 12 ms one has to fix the window size to 7.5 kBytes (corresponding to 5 full size Ethernet packet).

	description
Client nodes	Sun Fire V20z (bi-opteron)
Kernel	GNU/Linux 2.6.18.3
TCP variant	Bic, with SACK
iperf version	2.0.2
Topology	Butterfly
Bottleneck	1 Gb/s
RTT	12 ms
Sources nb.	100
Source rate	5 Mb/s
Exp. duration	8 hours
Flows nb.	5.10^6
Aggregation time	$\Delta = 100$ μs
Flow timeout	`timeout`= 100 ms

Table 1: Experimental global parameters.

4.4 Experiments description

Using the facilities offered by Grid5000, our metrology facilities and the rate control mechanisms for traffic generation, several experiments were performed, and we elaborate here on their rationale. First, the general experimental conditions are presented.

The primary interest here is the effect of flow size distributions on self-similarity, when each client behaves like a ON/OFF source model, where a ON period corresponds to a flow emission, and a OFF period to a silent source. The ON (respectively the OFF) lengths are random variables drawn independently and identically distributed, following the specific probability distribution P_{ON} (respectively P_{OFF}) we want to impose on the flows duration $\tau_{\rm ON}$ (respectively, on the silent periods, $\tau_{\rm OFF}$). The emission of packets in each flow is controlled by one of the methods described in the previous Section, each source rate being limited to 5 Mb/s to avoid congestion at the 1 Gb/s bottleneck.

All experiments consist of one trace of 8-hours traffic generation, representing a total of approximately $n = 5.10^6$ flows. As explained before, flows are reconstructed from the traces and we extract their flow sizes (in packets) $W = \{w_i,\, i = 1 : n\}$. Grouping and counting the packets in each contiguous time interval of width $\Delta = 100$ μ s, yields the aggregated traffic time series $X^{(\Delta)}(t)$.

In order to clearly define the terms of application of the Taqqu's Theorem on real traffic traces, as well as to identify possible interactions with other factors, we designed four series of experiments whose parameters are summarized in table 2.

Experiment A. This is the cornerstone experiment to check relation (14). Distribution of the ON periods are prescribed to Pareto laws with mean $\mu^{\rm ON} = 0.24$ s (corresponding to a mean flow size of $\langle P \rangle = 100$ packets). The experiment is repeated ten times with different prescribed tail index $\alpha_{\rm ON}$, varying from 1.1 to 4. OFF periods are kept exponentially distributed with mean $\mu^{\rm OFF} = \mu^{\rm ON}$. For each value of $\alpha_{\rm ON}$, an experimental point $(\widehat{\alpha}_{\rm ON}, \widehat{H})$ is empirically estimated.

	Proto	Band lim	α_{ON}	α_{OFF}	$\langle P \rangle$	meas param
A	TCP	PSP HTB TCP	1.1–4	-	100	$\widehat{H}$ $\widehat{\alpha}$
	UDP	iperf				
B	TCP	TCP	1.1 – 4	-	100 1000	$\Delta_{j_1}^{\langle P \rangle}$
C	TCP	TCP	1.5	1.1	100	$\widehat{H}_{\text{LRD}}$
			1.1	1.5		
D	TCP	TCP	1.1 – 4	-	100	$\widehat{h}_{\text{loc}}$
	UDP	iperf				

Table 2: Experimental conditions summary.

Moreover, to evaluate the possible influence of the protocol, and of the workload generation mechanism, the same series of experiments is reproduced with TCP (window size limitation) and with UDP (user-level packet pacing) first, and then using PSP, HTB and TCP throughput controls. The exact same trial of random variables defining the flow lengths is used for all experiments that imply the same probability law P_{ON}.

Experiment B. Under similar conditions as in series A, the mean of the ON periods takes on two different values $\mu^{\text{ON}} = 0.24$ s and $\mu^{\text{ON}} = 2.4$ s, corresponding to mean flow sizes $\langle P \rangle = 100$ and $\langle P \rangle = 1000$ packets, respectively. The objective is here to relate $\langle P \rangle$ to the lower scale bound $\Delta_{j_1} = 2^{j_1}\Delta$ defining a sensible regression range to estimate H.

Experiment C. The protocol (TCP), the throughput limitation mechanism (TCP) and the mean flow size ($\langle P \rangle = 100$) being fixed, we investigate now to role of the OFF periods distribution on the self-similar exponent H. In both experiment, P_{ON} and P_{OFF} are set to Pareto HT distributions with means $\mu^{\text{ON}} = \mu^{\text{OFF}} = 0.24$ s. In a first case, the tail index $\alpha_{\text{ON}} > \alpha_{\text{OFF}}$, and conversely $\alpha_{\text{OFF}} > \alpha_{\text{ON}}$ in a second case.

Experiment D. The last series of experiments aims at investigating self-similarity at finer scales (lower than the RTT scale), and whose origin is distinct from LRD phenomena. The changing paremeter is the tail index as in experiment A, yet the scaling law index will be estimated in the short time-scales limit, in order to characterize the traffic burstiness from the process $X^{(\Delta)}$. Under the same experimental conditions as in series A, we then evaluate how the protocols (TCP versus UDP) entail a significant change in the traffic burstiness.

5 Results and discussion

5.1 Verifying the Taqqu's relation

For every traces, we use the wavelet-based methodologies described in Section 3 for heavy-tail and self-similarity analyses. The estimated tail index $\hat{\alpha}$ results from the linear regression of Eq. (11) applied to the flow size sequence W, where

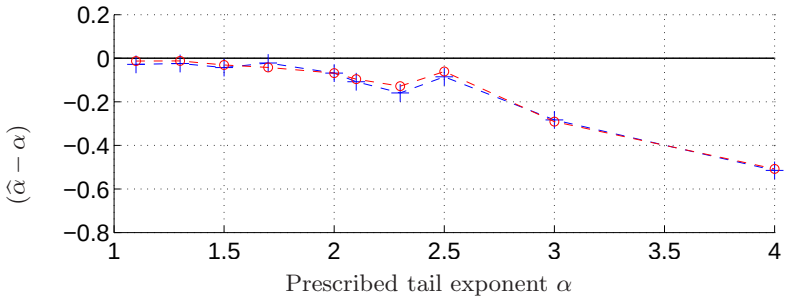

Figure 4: Difference between the prescribed HT index α and the actually estimated HT index $\widehat{\alpha}$ for the set of different values of α used in experiment series A (see Tab. 2) for two different protocols: TCP (+) and UDP (∘).

a sixth order derivative of a Gaussian wavelet is systematically used. The self-similarity index $\hat{H}$ is estimated from the LD plots of the aggregated time series $X^{(\Delta)}$, using a standard Daubechies wavelet with 3 vanishing moments [26].

5.1.1 Tail index estimates

Proceeding with experiment A, for different values of the tail index of flow size distribution, Fig. 4 displays the differences between the prescribed value α and the actually estimated value $\widehat{\alpha}$. The two experimental curves, corresponding to TCP and to UDP protocols respectively, superimpose almost perfectly. Beyond coherence with the fact that the exact same trial of random variables defining the flow lengths is used in both cases, such a concordance demonstrates that the flow reconstruction procedure, for both TCP or UDP packets grouping, is fully operative, notably including a relevant `timeout` adjustment (`timeout` = 100 ms).

Fig. 4 also shows an increasing difference $(\widehat{\alpha} - \alpha)$ with α. In our understanding, this is not caused by an increasing bias of the HT estimator, which is known to perform equally well for all α values. It is rather caused by the natural difficulty to prescribe large values of α over fixed duration. Indeed, as α increases, large flows become more rare, and the number of observed elephants during the constant duration (8 hours) of the experiments naturally decreases, then deviating from a statistically relevant sample. This observation is fully consistent with arguments developed in [32]. Notwithstanding this satisfactory agreement, in the sequel we will systematically refer to $\widehat{\alpha}$ rather than to the prescribed α.

5.1.2 LD-description

Fig. 5 shows typical LDs of aggregated traffic time series, obtained under similar conditions (experiment series A of Tab. 2, TCP protocol), for 4 different values of α. Such plots enable a generic phenomenological description of LDs: 3 different range of scales can be visually idendify, whose bounds do not seem to drastically vary with α:

Coarse scales: In the coarse scale domain, a clear scaling behavior is systematically observed. As mentioned earlier, Taqqu's Theorem relates heavy tails

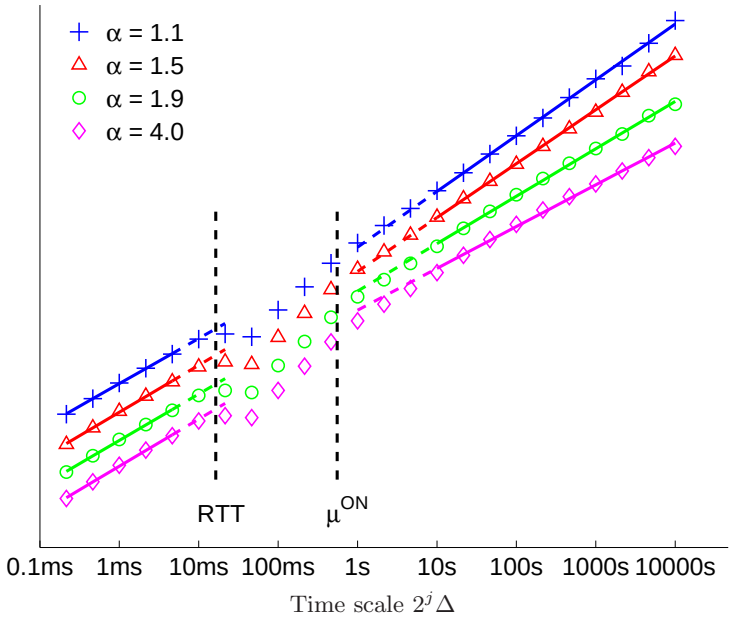

Figure 5: Wavelet log-diagrams $\log S(j)$ versus time scale j of aggregated traffic (aggregation interval $\Delta = 100\mu$ s). Log-diagrams correspond to 4 time series obtained under similar experimental conditions with the protocol TCP, with 4 different values of α: 1.1 (+), 1.5 ($\triangle$), 1.9 ($\circ$) and 4.0 ($\Diamond$).

and self-similarity in the asymptotic limit of coarse scales. Therefore, the coarse scale scaling exponent, denoted $\widehat{H}$, is a candidate to match that involved in Relation (14).

Fine scales: At fine scales, another clear scaling behavior is also observed. However, the corresponding fine scale scaling index, denoted h, is no longer related to Taqqu's Theorem prediction but rather to a local regularity property of the data.

Medium scales: Intermediate scales mostly connect the two scaling behaviors happening for fine and coarse scales, but exhibit no particular generic shape.

In Fig. 5, vertical lines materialize the two transitions scales between the three depicted domains and can hence be identified as characteristic time scales of the data. Let us now investigate the nature of these characteristic times.

5.1.3 Coarse scales domain lower bound

It is alluded in [22] that the range of scales where self-similarity can be measured is beyond a characteristic scale, referred to as the *knee* of the LD, and that it is essentially controlled by the mean flow duration. To investigate this argument in the context of our analyses, we designed two experiments series with two different values of the mean flow duration (series B of Tab. 2). For each case, all the LDs corresponding to the different values of α are computed. To emphasize the impact of the mean flow durations, we substracted to each LD, the asymptotic linear trend, obtained by linear regression between a scale

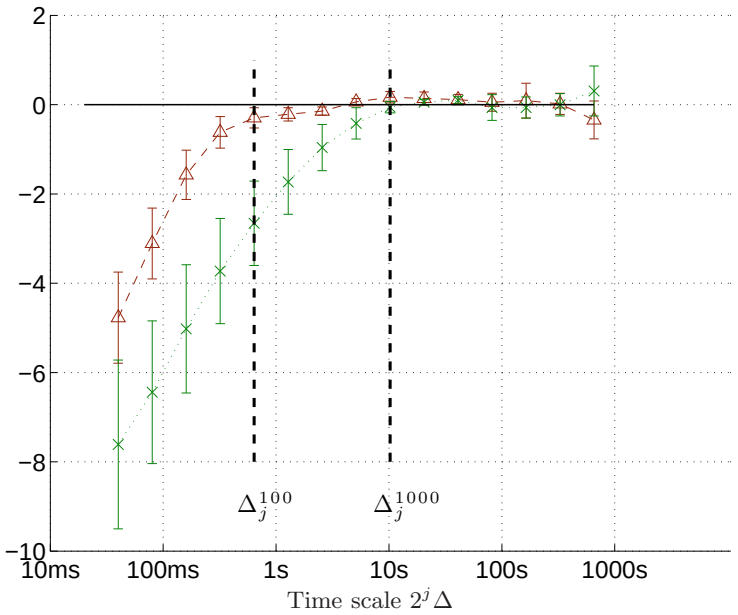

Figure 6: Averaged normalized log-diagrams for two different mean sizes $\langle P\rangle$ of the transmitted flows: (×) $\langle P\rangle = 1000$ packets – (△) $\langle P\rangle = 100$ packets .

Δ_{j_1}, clearly above the *knee* position, and the maximum available scale $\Delta_{j_{\max}}$. Fig. 6 shows, both for $\langle P\rangle = 100$ and $\langle P\rangle = 1000$ the mean and standard deviation over all normalized LDs. A slope break can be clearly seen on each graph: at scale $\Delta_j^{100} = 0.64$ s when $\langle P\rangle = 100$ and at scale $\Delta_j^{1000} = 10.28$ s when $\langle P\rangle = 1000$. Although for $\langle P\rangle = 100$, the knee effect slightly smoothes out, the linear behavior observed for $\langle P\rangle = 1000$ clearly extends with the same slope beyond Δ_j^{1000} up to Δ_j^{100}. Unquestionably, the measured knee position undergoes the same variations as the mean flow duration, both quantities being in the same order of magnitude: $\Delta_j^{100} \simeq \mu_{100}^{\mathrm{ON}}(0.24\text{ s})$ and $\Delta_j^{1000} \simeq \mu_{1000}^{\mathrm{ON}}(2.4\text{ s})$. This analysis confirms the intuition that the coarse scale range, where self-similarity is to be measured, lies above the *knee* of the LD, whose position is in the same order of magnitude as the mean flow duration. The coarse scales can then be termed: the *flow scales*, or the *file scales*.

5.1.4 Protocol, rate limitation and coarse scales

As we investigate Taqqu's relation, we now focus on the coarse scale domain. To inquire on the impact of the protocol on the coarse scales, Fig. 7 shows the LDs obtained with two different protocols : TCP and UDP (for $\alpha = 1.5$). Fig. 7 evidences the central feature that both LDs are undistinguishable in the coarse scale domain. We conclude that protocol has no impact on the coarse scale SS, at least in our experimental conditions, where source rate limitation precludes congestion.

Similarly, to inquire on the impact of the rate limitation mechanism on the coarse scales, Fig. 8 shows typical LDs ($\alpha = 1.5$, TCP) obtained with three different rate limitation mechanisms: PSP, HTB and TCP window limitation.

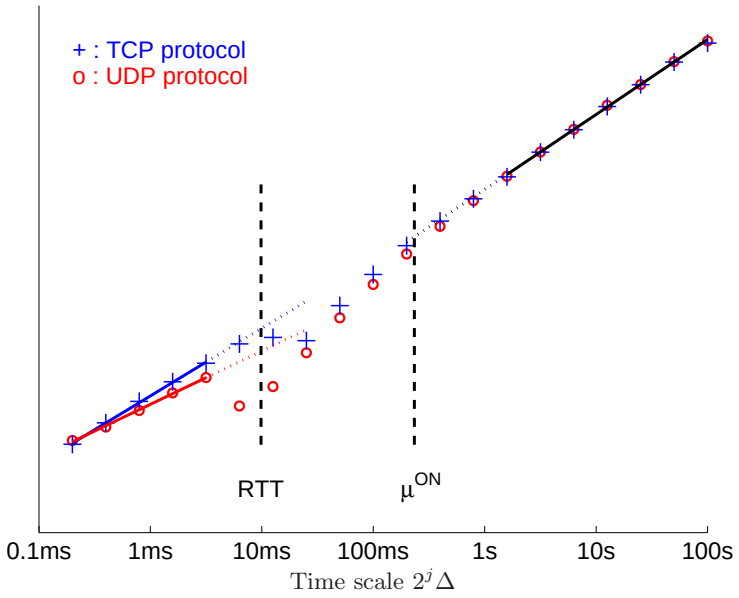

Figure 7: Wavelet log-diagrams $\log S(j)$ versus time scale j of aggregated traffic (aggregation interval $\Delta = 100\mu s$). Log-diagrams correspond to two time series obtained under similar experimental conditions, for $\alpha = 1.5$, with two different protocols: TCP (+) and UDP ($\circ$).

As the three LDs cannot be distinguished one from the other in the coarse scale domain, we conclude that the rate limitation mechanism has no impact on the scaling at coarse scales.

5.1.5 H versus α

Practically, to perform an empirical validation of Eq. 14, we need to estimate the scaling parameter H and thus to carefully choose the range of scales where the regression is to be performed. Although the *knee* position has been related to a measurable experimental parameter (the mean flow duration), a systematic choice of the regression range at coarse scales would certainly be hazardous. Instead, we defined for each trace an adapted regression range, based on a linearity criterion, and find that all regression ranges defined like this, encompass a scale interval ($\max_\alpha \Delta_{j_1} = 20.5$ s and $\Delta_{j_{max}} = 1310$ s), significantly extended to warrant statistically reliable SS exponent estimates.

Fig. 9 plots the estimates of coarse scale SS exponents against those of the HT indices. Confidence intervals for $\widehat{H}$ displayed on the graphs are supplied by the estimation procedure detailed in Section 3.1.2 [7, 6]. Such estimations are conducted independently for TCP and UDP protocols. For both protocols, estimations show a very satisfactory agreement with Taqqu's Theorem prediction. To the best of our knowledge, this theoretical relation between self-similarity and heavy tail had never been observed with such a satisfactory accuracy, (over a large and significant range of α values). For instance, and although no definitive interpretation has been proposed yet, the offset below the theoretical relation for α close to 1, and the offset above the horizontal line for α larger than 2 have

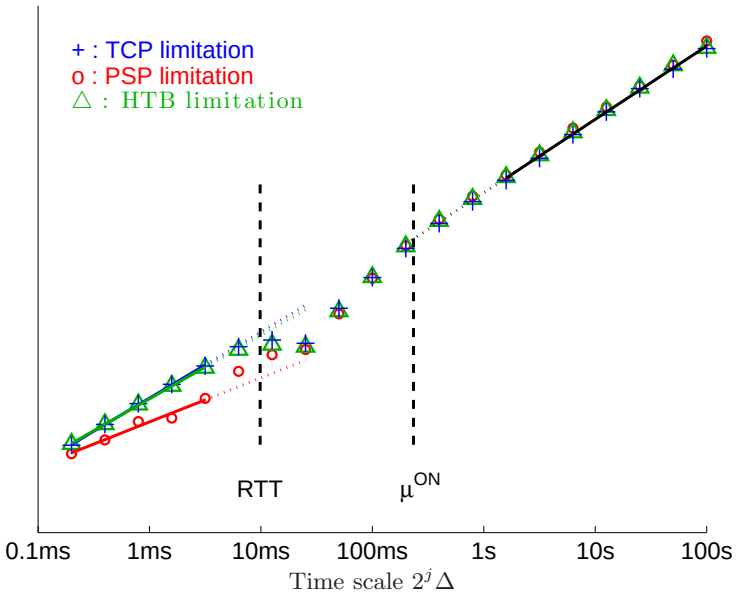

Figure 8: Wavelet log-diagrams $\log S(j)$ versus time scale j of aggregated traffic (aggregation interval $\Delta = 100\mu s$). Log-diagrams correspond to 3 time series obtained under similar experimental conditions, for $\alpha = 1.5$, with three different rate limitation mechanisms : TCP (+), PSP (∘) and HTB (△).

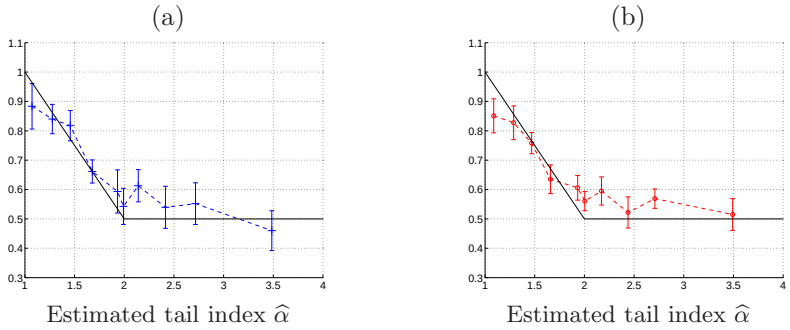

Figure 9: Estimated Self-Similar index $\widehat{H}$ of the aggregated traffic (aggregation interval $\Delta = 10\mu s$) versus estimated tail index $\widehat{\alpha}$ of the corresponding flow size distribution. Solid plots represent the theoretical model of relation (14), dashed plots correspond to experimental results: (a) with TCP protocol ; (b) with UDP protocol.

been drastically reduced when compared to similar analyses results reported in the literature (cf. e.g., [28]). This accuracy results from a number of factors: First, the statistical tools for estimating H and α are chosen amongst the most recent and robust (notably the proposed estimator for α had never been applied before to Internet data) ; Second, the asymptotic coarse scale nature of Taqqu's Theorem is really accounted for by performing estimation in the limit of really

α_{ON}	α_{OFF}	$\widehat{\alpha}_{\text{ON}}$	$\widehat{\alpha}_{\text{OFF}}$	$\widehat{H}$
1.5	1.1	1.502	1.113	0.869±0.062
1.1	1.5	1.080	1.505	0.865±0.055

Table 3: HT distributed OFF periods

coarse scales ; Third, this is made possible thanks to the use of really long duration, stationary and controlled traffic time series, wich is enabled by the use of Grid5000 platform.

Additionally, our analyses do confirm that TCP and UDP protocols do not impact this relation, at least in conditions corresponding to our experimental setup, where congestion is avoided by source rate limitation. This is in clear agreement with the findings reported in [19], or [21], showing that TCP is not responsible for the observed self-similarity. However, despite these earlier results, a non negligible number of contributions debated, investigated and argued in favor of an impact of protocols on self-similarity. Our analyses clearly show that the range of scales where protocols impact the LD is far below the characteristic time scales involved in self-similar phenomena.

As long as we actually consider coarse scales (larger than the mean duration of a flow), the only cause for self-similarity is the heavy tail in the flow sizes distribution.

5.1.6 OFF periods

To complement the experimental study of Taqqu's Theorem, the experiments of series C (see tab. 2) were designed to assess the influence of heavy tailed OFF periods on the coarse scale SS exponent H. Under experimental conditions detailed in Tab. 2, Tab. 3 sumarizes the resulting estimated coarse scale SS exponents.

For both experiments, the estimated value of the coarse scale SS index is about 0.86, which by inspection of Fig. 9 shows a satisfactory agreement with a tail index close to 1.1. We can conclude that, as predicted by the theory, if both ON-times and OFF-times are HT distributed, it is the smaller tail index among α_{ON} and α_{OFF}, that imposes and controls the coarse scale SS in traffic.

5.2 Further analyses of the LD

In the previous section, we focused on the coarse scales of LDs. Let us now turn to the medium and fine scales and study the influence of protocols and rate limitation mechanisms.

5.2.1 Medium scales

Firstly, let us notice that, while the mean flow duration gives an upper bound for the medium scale domain, RTT (12 ms) seems to correspond to its lower bound. Therefore, this medium scale range will be referred to as the *RTT-scales*. Although no scaling behavior is visible in this medium scale range, Fig. 7 shows a significant difference between the LDs obtained from TCP and UDP traffic. This is an expected result as RTT is the characteristic time of action of the TCP protocol.

Fig. 8 shows that there is no significant difference in this domain between the LDs corresponding to the three different rate limitation mechanisms. The characteristic time of action of the rate limitation is the mean inter-packet time. Due to the source rate limitation at 5 Mb/s achieved with 1500-Bytes packets, the mean inter-packet time for one source is 2.4 ms. As the mean number of sources emitting simultaneously is 50, the mean inter-packet time is 48 ms, which is much lower than RTT. Accordingly, the rate limitation does not impact the traffic at RTT scales.

5.2.2 Fine scales

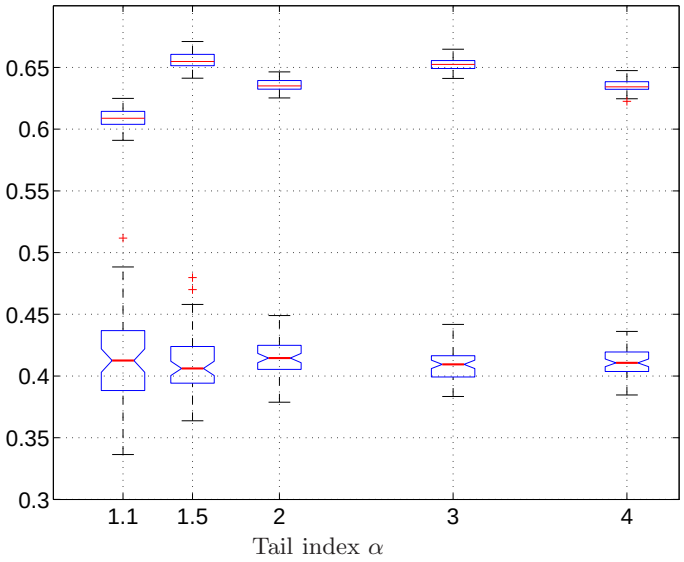

Figure 10: Fine scale scaling exponent h estimates on aggregated traffic time series ($\Delta = 100\mu s$). For different values of the tail index α governing the flow size distributions, h is estimated by linear regression of Log-diagrams (see Fig. 7) over the scale range $[0.2 - 5]$ms. Notched box-plots correspond to UDP protocol, regular box-plots to TCP protocol.

TCP and UDP impact on fine scales scaling. Figure 7 shows a good scaling behavior at fine scales, with a scaling index which seems to be different for UDP and TCP

To analyze in more details the fine scales scaling exponent, each 8-hours trace corresponding to a particular value of α (see experimental conditions of Experiment A in Table 2) are chopped into 66 short-length of duration $T = 100s$ each. The resulting time series are then analyzed independently and a fine scaling exponent h estimated. Hence, based on these 66 values of $\widehat{h}$, box-plots are displayed on Figure 10 for each theoretical value of α. TCP values remain roughly constant around $h \simeq 0.63$. Likewise for UDP, h does not seem to depend on α, but situates around 0.4, a significantly smaller value than that for TCP.

As these fine scales are smaller than RTT, and correspond to the *packet-scales*. Thus, the scaling index at these scales is sensitive to the packet sending

mechanism. When using UDP, packets are emitted individually, separated by an inter-packet interval (2.4 ms) imposed by iperf to respect the rate limitation (5 Mb/s). Then, UDP traffic is constantly and erratically varying. When using TCP, packets are sent by bursts containing up to 5 packets. Then TCP traffic is bursty, but also sparse, with "long" periods with no packets. We believe that this packet sending scheme difference, in close relationship with our experimental condition (source rate limitation used to avoid congestion) is sole responsible for the observed difference between TCP and UDP on the local regularity.

Bandwidth limitation impact on fine scales scaling. Figure 8 shows that the fine scale scaling index is approximately the same with TCP and HTB limitation, but it is very different with PSP limitation. This difference can again be explained by the packet sending mechanism. When using HTB limitation, packets are sent by bursts, in the same way as with TCP limitation. This explains why the local regularity observed with HTB is the same as the one observed with TCP limitation. On the contrary, when using PSP, packets are sent individually, in the same way as with UDP. Then, the fine scales scaling index observed with PSP is lower than the one observed with TCP limitation, as was the one observed with UDP.

6 Conclusions and perspectives

In this paper, we have revisited the relationship between file sizes and the self-similarity of traffic observed at link level. This work is based on three important innovative factors : the use of accurate estimation tools, a deeper analysis of Taqqu's Theorem applicability conditions and the use of a large scale reconfigurable experimental facility. The wavelet based estimation procedure for H is known has a state-of-the-art tool, being one of the most reliable and robust (against non stationarities). It has been used with care. The α index estimation procedure used here, shown to outperform previous available techniques, had never been used before with Internet data. Widely reckoned, the deeply asymptotic nature of Taqqu's Theorem has been better accounted for by conducting estimations of the self-similarity parameter at really coarse scales (coarse being quantified as scales far beyond the system dynamic). This asymptotic limit requires to produce traffic with particularly long observation duration, yet stationary, and well controlled. The nation wide and fully reconfigurable Grid5000 instrument enables generation, control and monitoring of a large number of finely controlled transfer sessions with real transport protocol stacks, end-host mechanisms, network equipments and links. Given such real and very long traces, we have been able to demonstrate experimentally, and with an accuracy never achieved previously with real data nor with simulations, that Taqqu's Theorem and the relation between self-similarity and heavy-tailness can actually be observed. In particular, we obtained a significant agreement between theoretical and experimental values at the transition points, around $\alpha = 2$. This is of particular difficulty for it mixes issues of different kinds regarding estimation of both H and α .

Concerning the discussion about the relationship between transport protocols and self-similarity – which remained quite controversial after and despite Crovella et al.'s meaningful contributions [14] –, our observations confirm that

protocols, rate mechanisms or packet and flow-level control mechanisms do not impact the observed self-similarity. Our analyses show that this is mostly because the ranges of scales related to self-similarity are far coarser than those (fine and medium scale) associated to such mechanisms. Ranges of scales have been quantified in terms of RTT and mean flow durations. We plan to further investigate this issue by designing specific experiments on our unique experimental Grid5000 tool. We aim at systematically investigating long term traffic traces generated under various congestion and aggregation levels, heterogenous source rates, mixed source protocols, various RTTs, different bottleneck and buffer capacities and new high speed transport protocols variants. We expect this will contribute to a better undestanding and a finer prediction of the network traffic in current and future Internet as well as to a relevant design of future transport and network control mechanisms.

References

[1] Hierachical token bucket packet scheduler. http://luxik.cdi.cz/∼devik/qos/htb/.

[2] Iperf, NLANR/DAST project. http://dast.nlanr.net/Projects/Iperf/.

[3] Iproute2, the linux foundation. http://www.linux-foundation.org/en/Net:Iproute2/.

[4] Ipsumdump. http://www.cs.ucla.edu/∼kohler/ipsumdump/.

[5] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi, and D. Veitch. Multiscale network traffic analysis, modeling, and inference using wavelets, multifractals, and cascades. *IEEE Sig. Proc. Magazine*, 3(19):28–46, May 2002.

[6] P. Abry, P. Flandrin, M.S. Taqqu, and D. Veitch. Wavelets for the analysis, estimation and synthesis of scaling data. In Kihong Park and Walter Willinger, editors, *Self-Similar Network Traffic and Performance Evaluation*. John Wiley & Sons, Inc., 2000.

[7] P. Abry, P. Gonçalves, and P. Flandrin. Wavelets, spectrum analysis and $1/f$ processes. In A. Antoniadis and G. Oppenheim, editors, *Lecture Notes in Statistics: Wavelets and Statistics*, volume 103, pages 15–29, 1995.

[8] R. J. Adler, R. E. Feldman, and M. S. Taqqu. *A Practical Guide To Heavy Tails*. Chapman and Hall, New York, 1998.

[9] D. G. Andersen, H. Balakrishnan, M. F. Kaashoek, and R. Morris. Resilient overlay networks. In *Proc. of the 18th ACM SOSP*, Oct. 2001.

[10] C. Barakat, P. Thiran, G. Iannaccone, C. Diot, and P. Owezarski. A flow-based model for internet backbone traffic. In *SIGCOMM Internet Measurement Workshop*, pages 35–47, New York, NY, USA, 2002. ACM Press.

[11] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In *SIGCOMM Internet Measurement Workshop*, Marseille, France, November 2002.

[12] A. Bavier et al. Operating system support for planetary-scale network services. In *Proc. of the 1st Symposium on Network System Design and Implementation*, Mar. 2004.

[13] R. Bolze et al. Grid'5000: a large scale and highly reconfigurable experimental grid testbed. *Int. J. of High Performance Computing Applications*, 20(4):481–494, nov 2006.

[14] M. E. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Trans. on Networking*, 5(6):835–846, December 1997.

[15] M. E. Crovella, M. S. Taqqu, and A. Bestavros. Heavy-tailed probability distributions in the World Wide Web. In Robert J. Adler, Raisa E. Feldman, and Murad S. Taqqu, editors, *A Practical Guide To Heavy Tails*, chapter 1, pages 3–26. Chapman and Hall, New York, 1998.

[16] P. Doukhan, G. Oppenheim, and M.S. Taqqu. *Long-Range Dependence: Theory and Applications*. Birkhäuser, Boston, 2003.

[17] A. B. Downey. Evidence for long-tailed distributions in the internet. In *SIGCOMM Internet Measurement Workshop*, pages 229–241, New York, NY, USA, 2001. ACM Press.

[18] N. Duffield, C. Lund, and M. Thorup. Estimating flow distributions from sampled flow statistics. *IEEE/ACM Trans. on Networking*, 13(5):933–946, October 2005.

[19] D. R. Figueiredo, B. Liu, A. Feldmann, V. Misra, D. Towsley, and W. Willinger. On TCP and self-similar traffic. *Performance Evaluation*, 61(2-3):129–141, 2005.

[20] P. Gonçalves and R. Riedi. Diverging moments and parameter estimation. *J. of American Statistical Association*, 100(472):1382–1393, December 2005.

[21] L. Guo, M. Crovella, and I. Matta. Corrections to "How does TCP generate pseudo-self-similarity?". *SIGCOMM CCR*, 32(2), 2002.

[22] N. Hohn, D. Veitch, and P. Abry. Cluster processes, a natural language for network traffic. *IEEE Trans. on Sig. Proc. – Special Issue on Sig. Proc. in Networking*, 8(51):2229–2244, October 2003.

[23] T. Karagiannis, M. Molle, and M. Faloutsos. Long-range dependence - ten years of internet traffic modeling. *IEEE Internet Computing*, September 2004.

[24] Y. Kodama, T. Kudoh, T. Takano, H. Sato, O. Tatebe, and S. Sekiguchi. GNET-1: Gigabit ethernet network testbed. In *Proc. of the IEEE Int. Conf. Cluster 2004*, San Diego, California, USA, Sept. 20-23 2003.

[25] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic (extended version). *ACM/IEEE Trans. on Networking*, 2(1):1–15, February 1994.

[26] S. Mallat. *A Wavelet tour of signal processing*. Academic Press, 1999.

[27] J. H. McCulloch. Measuring tail thickness to estimate the stable index alpha: A critique. *American Statistical Association*, 15:74–81, 1997.

[28] K. Park, G. Kim, and M. Crovella. On the relationship between file sizes, transport protocols, and self-similar network traffic. In *Int. Conf. on Network Protocols*, page 171, Washington, DC, USA, 1996. IEEE Computer Society.

[29] K. Park and W. Willinger. *Self-Similar Network Traffic and Performance Evaluation*. John Wiley & Sons, Inc., New York, NY, USA, 2000.

[30] V. Paxson and S. Floyd. Wide area traffic: The failure of Poisson modeling. In *SIGCOMM*, pages 257–268, New York, NY, USA, 1994. ACM Press.

[31] S. I. Resnick, H. Dress, and L. De Haan. How to make a hill plot. *OR & IE University*, 1998.

[32] M. Roughan, J. Yates, and D. Veitch. The mystery of the missing scales: Pitfalls in the use of fractal renewal processes to simulate LRD processes. In *ASA-IMS Conf. on Applications of Heavy Tailed Distributions in Economics, Engineering and Statistics*, American University, Washington, DC, June 1999.

[33] S. Soudan, R. Guillier, and P. Vicat-Blanc Primet. End-host based mechanisms for implementing flow scheduling in gridnetworks. In *GridNets 2007*, Oct. 2007.

[34] R. Takano, T. Kudoh, Y. Kodama, M. Matsuda, H. Tezuka, and Y. Ishikawa. Design and evaluation of precise software pacing mechanisms for fast long-distance networks. In *PFLDnet*, Lyon, France, 2005.

[35] M. S. Taqqu, W. Willinger, and R. Sherman. Proof of a fundamental result in self-similar traffic modeling. *SIGCOMM CCR*, 27(2):5–23, 1997.

[36] B. White et al. An integrated experimental environment for distributed systems and networks. *ACM SIGOPS Operating Systems Review*, 36(SI):255–270, 2002.

[37] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through high-variability: statistical analysis of ethernet lan traffic at the source level. *IEEE/ACM Trans. on Networking*, 5(1):71–86, 1997.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

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