Contributions to the Analysis of Scaling Laws and Quality of Service in Networks: Experimental and Theoretical Aspects

Patrick Loiseau

Numéro d'ordre : 554
Numéro attribué par la bibliothèque : _ENSL554

THÈSE

en vue d'obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE LYON - ÉCOLE NORMALE SUPÉRIEURE DE LYON

spécialité : informatique
Laboratoire de l'Informatique du Parallélisme
École Doctorale de Mathématiques et Informatique Fondamentale

présentée et soutenue publiquement le 11 décembre 2009
par Monsieur **Patrick Loiseau**

Titre :

Contributions to the Analysis of Scaling Laws and Quality of Service in Networks: Experimental and Theoretical Aspects

Contributions à l'analyse des lois d'échelles et de la qualité de service dans les réseaux : aspects expérimentaux et théoriques

Directeurs de thèse :
Monsieur Paulo Gonçalves
Madame Pascale Vicat-Blanc Primet

Après avis de :
Monsieur Jean-Yves Le Boudec
Monsieur Rudolf Riedi
Monsieur Philippe Robert

Devant la commission d'examen formée de :
Monsieur Christophe Diot, membre
Monsieur Paulo Gonçalves, membre
Monsieur Daniel Kofman, membre
Monsieur Jean-Yves Le Boudec, membre/rapporteur
Monsieur Rudolf Riedi, membre/rapporteur
Monsieur Philippe Robert, membre/rapporteur
Madame Pascale Vicat-Blanc Primet, membre

Remerciements

Je tiens d'abord à remercier mes directeurs de thèse Paulo Gonçalves et Pascale Vicat-Blanc Primet, qui m'ont fait confiance pour venir participer au développement d'une activité de métrologie naissante dans l'équipe RESO. Malgré une différence de culture scientifique évidente, j'ai admiré l'ouverture d'esprit de Pascale, qui a toujours fait l'effort de comprendre mes travaux théoriques pour les guider au mieux dans leurs applications aux réseaux. Quant à Paulo, son soutien de tous les instants et les multiples discussions que nous avons eut m'ont permis d'apprécier sa rigueur et son intégrité scientifiques, dont j'espère avoir au moins partiellement hérité. J'espère m'être montré digne de la confiance, du temps et des efforts qu'ils ont investis en moi.

Je remercie Jean-Yves Le Boudec, Rudolf Riedi et Philippe Robert d'avoir accepté d'être rapporteurs de ma thèse. Leurs rapports ont été très précis et leurs remarques particulièrement pertinentes, et je ne peux être qu'admiratif devant les efforts qu'ils ont dû fournir pour analyser ma thèse avec un tel niveau de détail. Je remercie également Christophe Diot et Daniel Kofman d'avoir accepté d'être membres de mon jury de thèse, et de venir participer à ma soutenance. Leur recul extraordinaire sur le domaine des réseaux m'ont permis de mettre en perspective mon travail. Cela a été un très grand honneur pour moi que de telles personnalités scientifiques prennent la peine de juger avec sincérité mes travaux de thèse.

Je ne sais comment remercier Julien Barral, qui m'a initier avec intelligence et passion à la recherche en mathématiques. Il a eut la patience de m'enseigner des bases que mon cursus de physicien ne m'avait pas données. Puissé-je faire bon usage de son enseignement. Je remercie également Rolf Riedi, qui m'a accueilli chez lui, à Fribourg, pendant quelques jours. Je regrette que le travail que nous avions alors commencé n'ait pas abouti, mais j'espère bien avoir l'occasion de collaborer avec lui de nouveau dans l'avenir.

Je voudrais remercier chaleureusement Patrice Abry, Pierre Borgnat et Guillaume Dewaele qui ont partagé avec moi leur expérience sur l'analyse du trafic à mes débuts. Je regrette que nous n'ayons pas plus interagi dans la deuxième partie de ma thèse, mais leurs encouragements à la fin m'ont beaucoup touché. Je remercie enfin Stéphane Girard et Florence Forbes d'avoir partagé avec moi leur expertise sur l'estimation statistique.

La partie technique de mon travail a été constamment soutenue par des ingénieurs (Yichun Lin, Damien Ancelin et Matthieu Imbert) qui ont fait face avec abnégation à mes requêtes changeantes. Qu'ils soient ici remerciés à la hauteur de leurs efforts.

Je remercie les nombreux chercheurs qui ne sont pas nommément mentionnés ci-dessus mais avec qui j'ai eut des discussions scientifiques très enrichissantes qui ont contribuées à ma formation. J'ai également apprécié l'interaction avec mes collègues de l'équipe RESO, du laboratoire d'informatique et de l'ENS Lyon.

Je remercie enfin mes amis et ma famille qui ont subi les moments difficiles qu'implique nécessairement la rédaction d'une thèse, tout en contribuant à les limiter. Grâce à eux, je garderai un excellent souvenir de mon passage à l'ENS Lyon.

Contents

Résumé en français

Dans un contexte d'expansion rapide de l'Internet, la caractérisation statistique du trafic dans les réseaux est une préoccupation centrale pour les fournisseurs d'accès, autant que pour les équipementiers et les développeurs de nouveaux protocoles. En effet, une bonne compréhension du trafic dans les situations représentatives des comportements des utilisateurs est une clé essentielle pour pouvoir garantir les performances au sens large, regroupées sous le terme de *qualité de service*.

Une avancée majeure dans cette direction a été la découverte en 1993 par Leland et al. et par Paxson et Floyd du caractère auto-similaire à grande échelle des séries temporelles de trafic agrégé dans le cœur des réseaux; suivie en 1997 du modèle de Taqqu et al. qui explique la présence de ces lois d'échelle par l'agrégation d'un grand nombre de sources ON/OFF dont les périodes ON (ou OFF) suivent une distribution à queue lourde. Ces résultats ont marqué le début d'une forte activité de recherche sur la modélisation des lois d'échelle dans les réseaux et leur impact sur la qualité de service.

Cependant, ces modèles théoriques se basent sur des hypothèses simplificatrices, inévitables pour en assurer la solvabilité, mais qui en limitent l'applicabilité aux réseaux réels. En effet, ils considèrent un modèle où l'émission par les sources est décorrélée du reste du système. Si ce modèle d'émission représente bien le comportement du protocole UDP, il s'avère être, dans une large mesure, inadapté à la description du comportement du protocole TCP (utilisé à plus de 80% dans l'Internet actuel), qui lui, est basé sur un mécanisme de contrôle liant l'émission des sources à la dynamique du système.

La complexité introduite par ce mécanisme de contrôle fait immédiatement ressortir la nécessité d'une étude expérimentale des caractéristiques du trafic. Cependant, une telle étude nécessite d'être capable de générer du trafic depuis un grand nombre de sources parfaitement contrôlées et de mesurer le trafic qui en résulte. C'est un problème très ardu si l'on veut utiliser un système réaliste comportant des liens à très haut débit (jusqu'à 10 Gbps), particulièrement lorsqu'il s'agit de mesurer le trafic à la fois en entrée et en sortie d'un buffer pour en étudier la dynamique. Pour cette raison, les études empiriques se sont souvent restreintes à l'utilisation de simulateurs et de nombreuses questions restent ouvertes, comme celles (fondamentales) de l'impact des lois d'échelles (ou plus généralement de la distribution des tailles de flux) sur la qualité de service ou encore de la modélisation du trafic prenant en compte la corrélation des sources introduite par leur interaction au niveau du buffer (en TCP).

Dans mon travail de thèse, j'ai choisi de m'intéresser aux lois d'échelles dans les réseaux et à leur impact sur la qualité de service, tant du point de vue expérimental que théorique et mes contributions ont été les suivantes.

Développement d'une plate-forme expérimentale à très haut débit pour la métrologie à grain fin. [4] [10, 9] [12]

Comme pré-requis pour toute tentative de modélisation du trafic, il est nécessaire de posséder une plate-forme expérimentale capable de générer du trafic depuis un grand nombre de sources, et de pouvoir mesurer le trafic agrégé. Pour cela, j'ai participé au développement d'une telle plate-forme, basée sur l'utilisation du testbed entièrement contrôlable Grid'5000 pour la génération du trafic. J'ai également participé au développement de l'outil *MetroFlux* qui permet de capturer sans perte le trafic traversant un lien à très haut débit (jusqu'à 10 Gbps). Malgré les difficultés liées à l'obtention d'une telle plate-forme fiable, nous sommes parvenus à mettre au point un dispositif expérimental unique qui permet de générer du trafic contrôlé depuis un grand nombre de sources et de monitorer le buffer commun à ces sources pour en étudier la dynamique.

Etude de l'auto-similarité à grande échelle dans le trafic agrégé. [1] [6, 7] [8, 11]

La première étude expérimentale que j'ai menée a visé à valider expérimentalement le modèle de Taqqu et al. qui lie le paramètre de Hurst du trafic agrégé H à l'indice de queue lourde de la distribution de la taille des flux α. Ce modèle n'avait jusqu'alors été validé que partiellement dans deux situations: soit sur des traces de trafic réel venant de l'Internet où l'on ne peut alors observer qu'un point de fonctionnement de la relation; soit sur des simulateurs. Grâce à l'utilisation de longues expériences stationnaires parfaitement contrôlées, et à l'utilisation des estimateurs les plus performants connus à l'heure actuelle, j'ai pu obtenir des résultats bien meilleurs que ceux précédemment observés. Cela m'a aussi permis de montrer les limites de validité de ce modèle, ainsi que de montrer que du fait de son caractère "grande échelle", ce résultat reste valable en TCP en l'absence de congestion qui corrèle les sources.

J'ai ensuite montré que le résultat de Taqqu et al. reste valable en présence de pertes aléatoires modérées, mais que si le taux de perte devient important, le mécanisme de contrôle de TCP modifie trop profondément l'émission des paquets et oblige à reconsidérer la notion de flux. Puis, j'ai montré qu'en situation de congestion où le trafic agrégé sature à la capacité du lien, ce résultat reste valable en considérant le trafic d'un sous-ensemble des sources.

Enfin, j'ai étendu le modèle de Taqqu et al. pour prendre en compte la corrélation entre les débits d'émission des sources et les tailles de flux. J'ai utilisé une approche basée sur des points de Poisson dans le plan ; approche classique dans le cadre des processus multiplicatifs possédant des propriétés de multifractalité, mais peu utilisée dans le cadre de processus additifs à longue mémoire. Ce modèle m'a permis d'interpréter des valeurs de H observées sur du trafic web, dans une situation qui n'entrait pas dans le cadre du modèle de Taqqu et al.

Estimation de l'indice de queue lourde de la distribution des tailles de flux sous échantillonage. [5, 3]

Parce que la capture de tous les paquets qui traversent un lien peut s'avérer coûteuse, nous avons alors été amené à considérer une procédure d'estimation de α à partir d'un sous-échantillon du trafic. J'ai développé un estimateur basé sur le principe du maximum de vraisemblance, dont les performances sont meilleures que celles des estimateurs existants. Cela m'a de plus permis de ré-interpréter ces derniers comme des approximations

de l'estimateur que j'ai proposé.

Etude de la qualité de service avec des sources ON/OFF à queue lourde.
Nous avons alors abordé la question de l'impact du paramètre de queue lourde de la distribution de la taille des flux sur la qualité de service.

Dans un premier temps, nous avons étudié le cas du protocole UDP, pour lequel une étude numérique basée sur des simulations est suffisante. Ces simulations m'ont permis de confirmer, pour des buffers de taille finie, des résultats obtenues dans le cadre théorique de buffers infinis: la qualité de service (en terme d'occupation moyenne du buffer et de taux de perte) est systématiquement dégradée par des faibles valeurs de α (synonymes de queues très lourdes), mais ce uniquement pour des tailles de buffers suffisamment grandes. J'ai également montré que la taille de buffer critique au delà de laquelle ce résultat est vrai est différente pour les deux métriques de qualité de service, résultat que j'ai interprété par la nature statistique différente de ces deux quantités.

Nous nous sommes ensuite intéressé au cas du protocole TCP, introduisant une rétroaction qui rend impossible l'étude par simples simulations. J'ai donc à nouveau utilisé la plate-forme expérimentale unique dont nous disposons pour réaliser de nombreuses expériences impliquant un grand nombre de sources ON/OFF TCP avec des périodes ON à queue lourde d'indice variable. Cela m'a permis de montrer que contrairement à une intuition légitime basée sur le cas du protocole UDP (et à des travaux passés sur des simulateurs), certains paramètres de qualité de service (le taux de perte en particulier) ne sont pas nécessairement dégradés par de faibles valeurs de α lorsque le protocole utilisé est TCP.

Identification de nouvelles lois d'échelles multifractales dans le trafic à l'aide d'un modèle markovien de TCP.
Le caractère "grande échelle" des résultats théoriques précédemment mentionnés et leur faible impact sur la qualité de service observé nous ont naturellement amené à considérer la modélisation du trafic à une échelle plus proche de la dynamique du protocole TCP (l'échelle RTT). Une modélisation multifractale s'est alors avérée adaptée et j'ai développé un modèle basé sur une évolution markovienne du trafic de chaque source prenant en compte les corrélations introduites par l'interaction des sources au niveau du buffer, à l'échelle RTT. Ce travail s'est divisé en deux composantes. Une première composante théorique a consisté en la démonstration d'un théorème presque sûr de grandes déviations pour une seule réalisation d'un processus stationnaire ergodique possédant des propriétés de mélange et associé à un principe de grandes déviations traditionnel (*i.e.* impliquant des moyennes d'ensemble). L'application de ce théorème au cas des chaînes de Markov m'a ensuite permit d'exhiber de nouvelles lois d'échelles multifractales dans le trafic TCP. Ces lois d'échelles, intimement liées à la dynamique du protocole TCP et à ses mécanismes de contrôle, permettent une analyse multi-résolution de différents paramètres liés aux performances de TCP: le débit et la "fairness" (équité entre plusieurs sources). Elles permettent la prédiction de bornes (au sens probabiliste) sur les performances atteintes dans une situation données, et sont donc d'un intérêt direct pour la communauté réseaux.

Abstract

In today's context of rapid expansion of the Internet, statistical characterization of network traffic is a central preoccupation for Internet Service Providers, as well as for networks infrastructures and protocols designers. A deep understanding of traffic characteristics in situations corresponding to typical users behaviors is indeed essential to guaranty *Quality of Service*, *i.e.* performance in the broad sense.

A major breakthrough in that direction was the discovery in 1993 by Leland et al. and by Paxson and Floyd of the self-similar nature of aggregate traffic time series at large time scales. Following up this seminal discovery, the theoretical work by Taqqu and collaborators in 1997 proposed a plausible origin for these scaling laws by showing that they can be generated by a large number of ON/OFF sources with heavy-tailed ON (and OFF) distributions. These results were followed by a large research activity on the impact of scaling laws on Quality of Service.

However, as these theoretical models rely on necessary simplifying assumptions to ensure their tractability, their practical application to real networks situations is of limited use. In particular, they assume independence between the sources emission and the rest of the system. Realistic in the context of UDP protocol behavior, this emission model turns unable to account for the specificity of TCP protocols (used in more than 80% of the transfers), which rely on control mechanisms that inevitably correlate the sources emission to the system dynamic.

The complexity introduced by this control mechanisms gives a major importance to the experimental study of the traffic characteristics. However, performing such a study in realistic conditions that involve a large number of sources and very high speed links (up to 10 Gbps), is very challenging. In particular, full capture of the traffic at such high rates is an arduous task, especially when synchronized captures at the input and output of a buffer are needed to study its dynamic. For this reason, empirical studies have often been limited to the use of simulators, and many fundamental questions related to TCP traffic remain open: What is the impact of scaling laws (or more generally of the file size distribution) on the Quality of Service? How to take into account, in TCP traffic models, the sources correlation introduced by their interaction at the buffer level? etc.

In my PhD work, I chose to focus on scaling laws in network traffic and their impact on Quality of Service, both from an experimental and theoretical viewpoint. My contributions are the following:

Development of an experimental platform for fine grain metrology at very high speed. [4] [10, 9] [12]

Platform based on the *Grid'5000* tested for traffic generation; and on the FPGA based device *GtrcNet* for traffic capture.

Study of self-similarity at large time scales in aggregate network traffics. [1] [6, 7] [8, 11]

Experimental validation of Taqqu's theorem; study of the self-similarity in lossy TCP traffics; extension to take into account the correlation between flow sizes and flow rates observed on real web traffic.

Estimation of the flow size distribution tail index from packet sampled data. [5,3]

Design of a Maximum Likelihood estimation and comparison with other existing estimators.

Study of the Quality of Service with heavy-tailed distributed ON/OFF sources.

Experimental study of several QoS metrics, with different tail index parameters, for both UDP and TCP protocols.

Identification of new multi-fractal scaling laws in TCP traffic, using Markovian models.

Theoretical study (proof of a large deviation theorem for one realization of a mixing process) and application to TCP traffic (experimental validation and applications).

Chapter 1

Introduction

1.1 Context

Communication networks

Communication networks are the support of the transmission of various types of information: voice, computer files, etc. They are constituted of nodes (machines, call servers, etc.) and links transporting the information between the nodes. Two main categories of communication networks can be distinguished. In *circuit switching* networks, used historically for telephony or currently in optical networks, an isolated *channel* is reserved between two nodes for the duration of the call. This way, once the connection is established, the transmission quality is guaranteed during the entire call. However, if no channel is available, a connection request might be blocked. The procedure which decides whether a request can be accepted or not is called *admission control*. On the other hand, in *packet switching* networks, consistent pieces of information (files for instance) are fragmented in *packets* emitted successively, that can take different routes to their destination and be multiplexed with packets from other sources. In such networks, even if a *logical connection* (*e.g.* a TCP connection) can be established between a source and a destination, there is no guarantee *a priori* on the links' availability, and subsequent packets might be lost or delayed. To limit packet losses, intermediate points of the network (routers and switches) can store a limited amount of packets in *buffers*. The well-known Internet is based on this packet-switching principle. It has greatly expanded over the last decade, and is still expanding fast (the french institute ARCEP[1] reports a 12% increase of the number of high-speed internet subscribers between June 2008 and June 2009 in France, reaching more than 18 million subscribers).

Quality of Service in packet networks

In this context, *Quality of Service (QoS)* is a central preoccupation for network operators. Since the "quality" of a connection cannot be guaranteed in packet-switching networks as it is in circuit-switching networks, the "service" provided (and eventually sold) by the operator has to be defined, and how its quality is quantified is not clear. Network QoS can be defined in many different ways and include a variety of concepts such as performance, security, and transmission reliability. We focus on the performance aspects, and typical QoS metrics are then delay (time taken for a packet or a flow to go from source to destination), jitter (variation of the delay), bandwidth (quantity of information sent per time unit, *i.e.* transmission speed) and loss rate. In this context, there are several ways to tackle the

[1] Autorité de Régulation des Communications Eléctroniques et des Postes, http://www.arcep.fr

question of QoS. We present only two approaches here, that radically differ in their basic principles: the first approach is deterministic and attempts to give users strict guarantees on given performance metrics. It is naturally based on the general principles of *resource reservation* and *admission control*. The second approach is probabilistic, and basically attempts to optimize the control of *best-effort* transfers based on some knowledge of the traffic distribution. Hybrid approaches, such as the *DiffServ* model [Blake et al., 1998] based on prioritization have also been proposed, but we do not present them here.

A deterministic approach to *guarantee* QoS. The deterministic approach consists in guaranteeing QoS to users, *e.g.*, a certain minimal bandwidth (the exact form of the performance guarantee is part of the Service Level Agreement (SLA) between the customer and the network operator). Since the operator's infrastructure is not easily and quickly extendable when the demand arrives, this is possible through reservation of some bandwidth during a fixed period of time, or admission-control procedures; and is equivalent to introducing circuit-switching properties in a packet-switching network. An example of such an approach is the *IntServ* model [Braden et al., 1994], which is based on admission control and resource reservation. The customer negotiates a service level at the connection, and if his requirement cannot be satisfied, the request is blocked. Guaranteeing performance between two nodes connected via a complex network is difficult. The elegant theory of *network calculus* [Le Boudec and Thiran, 2001] is able to provide deterministic bounds for several performance metrics like delay and bandwidth. However, the *IntServ* model has not been widely developed, mainly because of scalability issues (due to the "state-fulness" of this model), and of pricing and management issues. An alternative solution exploiting recent advances in dynamic control plans, has been recently proposed. It uses the idea of *reservation in advance* [Soudan, 2009]: a user reserves the previous day, *e.g.*, before 6 p.m., the amount of bandwidth he will need the following day and during what period of time. Obviously, admission control is also applied with this method, since, if the demand cannot be satisfied, it is refused. The chances for the demand to be satisfied increase if it is made long in advance. This creates an incentive for users to do so, which has the beneficial effect of helping the network operator manage his resources. To achieve the operator's goal (maximal profit), pricing schemes can be applied (e.g., reservations made long in advance are cheaper, strong constraints are more expensive), leading to complex optimization problems closely related to game theory, which are promising topics of current and future research. Such deterministic approaches to guarantee QoS are well-suited for users having bulk data to transfer and hard deadlines for their transmission. However, beyond the issue that the users are constrained to plan their transfers in advance; it has several scalability limitations which render it hard to deploy for all the traffic.

A probabilistic approach to *improve* best-effort transfers. The strict QoS guarantees offered by the deterministic approach do not correspond to the needs of many applications. Indeed, there has been much work to develop applications that are able to cope with network variability. For example, video applications implement large buffers on the receiver to absorb periods of network congestion when information does not arrive regularly, so that a strict bandwidth guarantee is not needed. The second approach that we consider is then somehow opposite: it does not rely on admission control or reservation policies to provide strict QoS guarantees. Instead the goal is to optimize the transfers and

resource usage in a best-effort context. User behavior is not modified in this approach. Instead, users are free to send whatever they want at any time, and it is the role of the network to provide reasonable performance. QoS is evaluated in probabilistic terms in this approach (for example, the mean delay is 100 ms or the delay is less than 100 ms 95% of the time) and is usually called *Quality of Experience (QoE)*. Evidently, QoS in this context strongly depends on the traffic matrix and distribution (*i.e.,* characterization of the users demand). However, given a traffic distribution, there are several ways to improve QoS in such an environment (without acting *directly* on user behavior). Firstly, the infrastructure (capacity of the links and size of the buffers) can be adapted to the demand. This is the *dimensioning* problem: for example, given the traffic distribution, determining what the optimal buffer size should be to achieve a small loss rate and not increase delay too much. Secondly, most of the Internet traffic relies on the TCP transport protocol. This protocol ensures reliability of the transmission by reemitting lost packets, but it also uses a congestion avoidance mechanism: if a packet is lost (sign of congestion), the emission rate is reduced, otherwise the emission rate is increased to exploit the remaining capacity. Optimization of this mechanism, to avoid loss events while maximizing resource utilization is a way to improve performance. Finally, manipulation of packets within a fixed size buffer is a promising approach: Active Queue Management (AQM) and packet scheduling are examples of such approach. AQM consists in detecting congestion before losses occur. Packets can then be dropped (with RED) or marked (with ECN) before the buffer is actually full, with the goal of limiting severe congestion events. Flow aware approaches, based on the knowledge of traffic and flows to achieve better forwarding of packets are also promising [Bonald et al., 2002].

Statistical characterization of network traffic

All the aforementioned probabilistic approaches to network QoS require a deep understanding of the properties of the traffic generated by users. In packet-switching networks, traffic consists of packets, and a full characterization naturally falls into the field of *point processes*: traffic is characterized by the arrival time of all the packets. If packets can have different sizes, then the process is said *marked*: characterization of each point include (in addition to its arrival time) its size. At the input points of the network's shared links, packets of several users are naturally interleaved: this is the equivalent of the *superimposition* operation in point processes theory. This operation is often called (source) aggregation and the corresponding router is the aggregation point. Due to the large number of users, and the naturally random human behavior, this aggregate traffic is well modeled by a random point process. This is an important property in the study of traffic from large-scale networks (like the Internet), that makes the use of probabilistic methods relevant to our work (random models might not apply well to, *e.g.*, traffic in a very small cluster of machines where deterministic effects of a particular application can prevail). There are various ways of characterizing a random point process. The finer way consists in describing the intervals between two consecutive points, by their marginal statistics (the probability that the interval between two consecutive packets has a certain value) and correlations (the conditional probability that the interval between two consecutive packets has a certain value, given the values of precedent intervals). However, one might not be interested in such a fine-grain description, and prefer a more coarse-grain description, based on the *count process*. This process is the number of packets in consecutive time windows of fixed size Δ. Consider-

ing this process corresponds to averaging the traffic at time scale Δ. Variations inside a time window are lost in this description; for example, one cannot distinguish the case where all the packets are at the beginning of the window from the case where packets are regularly spaced in the window. Validity of this description to tackle QoS problems is still justified by the existence of buffers which "absorb" small-scale variations of the traffic. Of course, careful choice of Δ is required to obtain accurate QoS predictions. The count-based description is then very practical, because it provides information at regularly spaced intervals (Δ). It can be viewed as a uniformly sampled continuous signal, for which many signal processing techniques are available to analyze statistical properties. A deep understanding of the statistical properties of this signal in situations corresponding to typical users' behavior is essential to understand and eventually improve QoS in networks.

Markovian modeling for phone networks. The statistical analysis of network communications started in 1917 with the works of Danish engineer Agner K. Erlang on phone networks. Erlang's approach is based on Markov chains: simple stochastic processes for which the probability of an event depends only on the preceding event. The Markovian approach of Erlang supposes that incoming calls arrive as a Poisson process (the simplest point process where inter-arrival times are independent and exponentially distributed). The corresponding count process in consecutive time windows has Poisson marginals and independent values. The number of customers in the system can then be described by a Markov chain if the call duration is exponentially distributed. This approach was proven to be justified in phone networks, and was used to dimension the network in order to ensure a very low blocking rate.

New statistical properties in packet-network traffic: scaling laws. At the beginning of packet-switching networks, the statistical characterization of traffic was based on Poissonian assumptions similar to the ones of Erlang's approach [Kleinrock, 1969]. However, after the congestion collapse of 1986, the AIMD mechanism was introduced in the TCP transport protocol [Jacobson, 1988]. The main reason of the extraordinary expansion of the Internet, both in term of size of the network and size of the transfered files, probably lies in this decentralized congestion control algorithm, which permitted for example the web success in 1992. In 1993, pioneering work by Leland et al., and by Paxson and Floyd showed that aggregate traffic of different kind of packet-switching networks possesses the *long-range dependence* property: the probability of an event depends not only on the previous event, but also on events that happened long in the past. The presence of the long-range dependence property invalidates Markovian modeling (unable to account for this type of long-term correlation), and consequently the dimensioning and control procedures based on this Markovian assumption. The long-range dependence property of a signal is often induced by self-similarity, an example of the more general concept of scaling laws, which express that no specific scale prevails in the system. Instead, different scales of the system are linked together by a *scale-invariant quantity* that remains invariant when calculated at different scales. In the case of self-similarity, this scale-invariant quantity characterizes the evolution of the variance (or, more generally, any finite order moment) of the signal across scales. It remains constant in the signal, and the scaling law is then called *monofractal*. A few years after the first observation of self-similarity in network traffic (in 1997), several works reported the presence of a different type of scaling laws:

multifractality at small time scales. The term *multifractality* denotes erratic variations of the scale-invariant quantity with time. Although it can refer to any scale-invariant quantity, it is usually understood to refer to the local regularity, and is then characteristic of a very high variability of the signal regularity. Scaling laws are of primary importance in the statistical characterization of network traffic, and will therefore be at the center of this thesis. They are valuable fingerprints of the system state and of the mechanisms underlying their presence. Moreover, they can have a strong impact on performance. They are, for example, sources of QoS degradation, and so need to be considered in design problems such as buffer sizing. A precise understanding of their origin is thus essential. A major breakthrough in this direction has been the theoretical work by Taqqu and collaborators in 1997. It posits that the heavy-tail nature of some probability distributions, mainly that of flow size distributions, suffices to generate traffic exhibiting long-range dependence. The idea that infinite variance distributions produce long-range dependence was already put forward by Mandelbrot (Taqqu's PhD advisor until 1972) since 1969 and throughout his work.

The need for an experimental study in realistic conditions.

The theoretical models previously mentioned and many other models that appeared later explaining the presence of self-similarity in network traffic and its impact on QoS are simplified to ensure their tractability. In particular, the source emission is assumed independent of the rest of the system. Although this represents well the UDP protocol's behavior[2], it is unable to account for the specificity of the TCP protocol, whose control mechanisms inevitably correlate the source emission to the system's dynamic. Due to their complexity, comprehensive modeling of all TCP specifications is very arduous and this gives a major importance to experimental studies of traffic characteristics. However, performing such a study in realistic conditions, implying a large number of sources and a very large spectrum of rates as in the Internet (up to 10 Gbps), is a very difficult task. In particular, full capture of the traffic (necessary to perform fine-grain analysis) at such high rates is an arduous challenge, especially when synchronized captures at the input and output of a buffer are needed to study its dynamic. For this reason, empirical studies have often been limited to the use of simulators. However, simulators have limitations of their own. Current simulators (like ns2) have computation-cost limitations that reduce their usability for large-scale high-speed networks. Moreover, such discrete event simulators introduce unavoidable determinism and cannot account for the natural randomness inherent to real systems. An experimental study in realistic conditions is then necessary to tackle the very important question:

Overall question: In realistic conditions, what are the statistical characteristics of network traffic? What is their impact on the Quality of Service in packet networks?

1.2 Presentation of the system

The single bottleneck system. To precise and answer this question, this thesis focuses on the following simplified system. A large number of sources share a single bottleneck

[2]Throughout this manuscript, we abusively equation "UDP traffic" to "non-controlled traffic", even though some UDP applications implement control mechanisms.

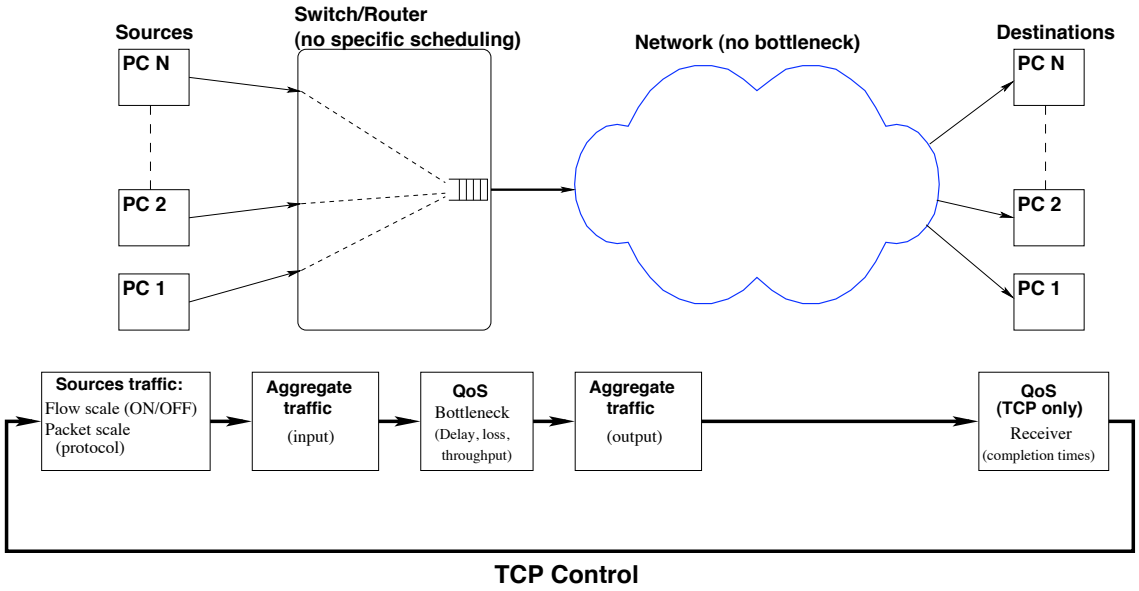

Figure 1.1: The single bottleneck system.

(Figure 1.1). The bottleneck applies no specific scheduling policy (First Come First Serve (FCFS)), and uses the simple *droptail* policy (packets are dropped only if no room is available in the buffer). No correlation is introduced *a priori* between the sources. In our work, the number of sources goes up to 100. This is only 1 or 2 orders of magnitude less than the typical access point of an Internet Service Provider (ISP), where the sources are the subscribers geographically grouped and sharing the access point. In a core point of the Internet backbone, several links carrying the traffic from multiple ISPs are aggregated and the number of users can be much greater, but the backbone links are usually over dimensioned so that QoS problems are much more critical at the access points. The general principles derived in our study, however, give interesting insights regarding the backbone situation as well. Obviously, such a single bottleneck system does not allow studying specific effects arising in networks of queues, like stability problems. We also do not tackle questions related to scheduling. However, this simple system is rich enough to permit a thorough study of the statistical characteristics of network traffic and in particular to identify the fundamental origins of scaling laws without introducing unnecessary complexity.

Description of user behavior. The statistical characteristics of aggregate traffic obviously depend on the characteristics of each source's emission, which we now describe. There are two levels of source characterization: the flow level and the packet level (or flow scale and packet scale). The flow level describes user behavior (what data is sent and when) at a coarse grain. To account for the typical user's behavior, we use the simple ON/OFF scheme: a source emits data during a certain ON period (called *flow time* or *flow duration*), then stops emitting for some OFF time usually called *idle time* or *think time*. Packets emitted during an ON period form a flow (a coherent set of related packets pertaining to a same entity, usually a file). The quantity of packets in a flow is called *flow length* or *flow size* and the quantity of bytes, *flow volume*. For a given flow size, the flow duration might vary, depending on the conditions. That is why our description of user

behavior is focused on flow sizes rather than ON times.

For a given source, flow sizes and OFF periods are assumed independent, and hence to complete the description, we only need to characterize their distributions. Based on recurrent observations on real Internet traces since 1996, we use heavy-tailed distributions, characterized by their *tail index* α. The packet-level description of the source traffic is governed by the transport protocol, which we now discuss.

The transport protocol. To complete the description of a source's emission, we need to describe how packets are emitted within a flow. This is controlled by the transport protocol. Two main protocols exist in the Internet that differ fundamentally in their basic principles: UDP and TCP. The UDP protocol is the simpler of the two. It emits regularly spaced packets, at some constant rate (number of packets per time unit). This emission mechanism is independent of the system's state (the rate does not change if packets are lost), and unreliable (lost packets are never recovered so that integral transmission of a file is not guaranteed). It can however be well suited for situations like voice or radiophony transmission, where partial information is sufficient. For radiophony broadcasting, for example, a certain percentage of information loss is acceptable without compromising the auditive impression. On the opposite, the TCP protocol is based on two principles proposed in 1973 by Cerf and Kahn: regulation and acknowledgment. The principle of acknowledgment is intended to permit reliable transmissions. After emission of a packet by a source, the destination acknowledges good reception of the packet (by sending an ACK packet). Non-received packets can then be retransmitted. The principle of regulation is intended to detect and avoid congestion in the network: if the network is congested, the source should slow down its emission rate, whereas if bandwidth is available it can emit more aggressively. A simple and efficient form of this regulation has been proposed in 1988 by Jacobson [Jacobson, 1988], on which is based the common TCP variant Reno that we present here. The source emits a *burst* with a certain number of packets called *congestion window*. After a certain time called Round Trip Time (RTT, basically corresponding to the transmission time to the destination and return), acknowledgments are received by the source. If a *single loss* is detected, the congestion window is divided by two, otherwise it is increased by one; the source then emits a burst of the corresponding value. This mechanism is called *Additive Increase Multiplicative Decrease* (AIMD). Single losses can be detected by the reception of acknowledgment packets from subsequent sent packets. In the case of *loss bursts*, where no acknowledgement is received during a certain time interval (a timeout called RTO), the congestion window is reset to the value one packet. The emission in the TCP case is then fundamentally different from the UDP case in several ways: firstly, it is *ack-clocked* (bursts of packets are emitted every RTT), and thus not regular. It introduces a new characteristic time scale: the RTT. More importantly maybe, the rate is adapted based on some *feedback*. It means that an open loop analysis of the system is not possible: *e.g.*, we cannot analyze the effect of the source emission on the buffer (*e.g.*, loss probability), without accounting for the reverse effect of the losses on the source emission. Finally, this feedback introduces correlation between the sources. Indeed, when one source reduces its sending rate because of a loss, other sources can benefit from this reduced load to increase their rate. In case of an ideal rate adaptation, this correlation takes a particular form: the total rate equals the capacity of the link. Practically though, non-instantaneous adaptation leads to periods of partially used link capacity. As the additive increase of

Jacobson's algorithm can be too slow at the beginning of a connection, until the first loss is detected or some threshold is reached, the *Slow Start* algorithm is used: the congestion window is doubled at each RTT. This permits to increase the sending rate faster at the beginning. After the first loss, the protocol enters the so called *Congestion Avoidance* phase, where Jacobson's algorithm is used in the simplest TCP variant, called Reno. The slow start mechanism is also used after a timeout, when the congestion window is reset to one. While the exponential increase during this phase corresponds to a fast increase, this justifies the term *Slow* Start, for a small number of packets (one) is first sent. Many other TCP variants than Reno have been proposed [Guillier, 2009], that change the evolution of the congestion window during the congestion avoidance phase. Basically, these variants try to minimize the time to increase the sending rate after a loss, using various growth shapes and levels of aggressivity. This presentation is oversimplified, and many other subtleties and diverse implementations of TCP mechanisms have appeared in the last two decades, that are partly responsible for the difficulty of building a comprehensive TCP model. The research community itself has limited confidence in complex models trying to account for all of these specifications and usually trust them less than experimental studies to give accurate results on the performance of TCP in various situations. Oversimplified models of TCP forgetting these subtleties and concentrating on essential mechanisms like AIMD, however, are able to provide qualitative insights and tendencies of major interest in both the understanding of TCP traffic behavior and the development of more efficient protocols.

The general questions in the single bottleneck system. The single bottleneck system that we study allows us to decompose the overall question of statistical characterization of the network traffic and QoS, by considering the different points of the network, represented on Figure 1.1. Basically, we want to characterize the traffic and QoS at each point of the network, and we articulate our work around the following general issues:

Aggregate network traffic characterization.

Source traffic characterization at flow scale.

Impact of the traffic characteristics on QoS (in performance terms).

TCP traffic at packet scale and TCP throughput predictability.

The approaches to these questions strongly depend on the used protocol (TCP, UDP). With UDP, because the source emission is independent of the system, the questions can be studied separately with an open-loop approach. For instance, based on the characteristics of the source traffic, we can determine the characteristics of the aggregate traffic at the input of the buffer; and knowing this input traffic, we can determine the QoS at the buffer (loss, delay, etc.), without coming back to the source emission. Characterization of the QoS at the receiver is irrelevant with UDP because the flow rate is almost constant and the completion time of a flow is thus proportional to its size. On the opposite, these questions cannot be studied separately with TCP, because of the feedback mechanism. Indeed, the characteristics of the source emission process influence the aggregate traffic shape, which itself has an impact on the QoS at buffer level (loss in particular), thus modifying the source emission process, etc. This is a closed-loop system, whose modeling is much more difficult than in UDP's case where fluid models can give satisfying results. An experimental approach is thus necessary, bringing the preliminary problem of possessing an appropriate tool:

Experimental platform reproducing realistic conditions.

All of these general questions have already received a major attention in various research communities such as mathematics and networks communities; both from the academic and the industrial side (for network operators in particular). Of course, they all have received partial answers in the literature, but some parts are still missing, in particular due to the difficulty of a large-scale experimental approach in realistic conditions. The next section details my contributions and the answers they bring to these questions; and relate the specific context in which my work has been done.

1.3 Summary of my contributions and context of my work

Development of an experimental platform for fine-grain metrology at very high speed. [4] [10, 9] [12]

As a pre-requisite to experimentally tackle the previously mentioned questions, it is necessary to possess a large-scale experimental platform, able to generate traffic from a large number of sources in realistic conditions (very high speed), and to measure the traffic going through a link. I participated in the development of such a platform, based on the fully controllable testbed Grid5000 for traffic generation. Our contributions included the development of *TranSim*, a tool to automate the generation of traffic from a large number of sources and the collect of logs and statistics for each experiment; and the development of *MetroFlux*, a tool to capture the traffic going through a link at up to 10 Gbps. I made these developments in collaboration with the INRIA engineers Yichun Lin (January 2007–June 2007), Damien Ancelin (April 2007–April 2009), and Matthieu Imbert (2008–). The tool *MetroFlux* is based on the FPGA device *GtrcNet* developed at AIST, Japan. At 1 Gbps, it is able to performed two synchronized captures, for example at the input and output of a buffer to study its dynamic. Despite the natural difficulty inherent to real equipments, we managed to obtain a reliable and powerful metrology platform, which I used throughout my thesis study network traffic and QoS.

Study of the self-similarity at large time scales in aggregate network traffic. [1] [6, 7] [8, 11]

This contribution gives answers to the question of the aggregate traffic characterization. It assumes knowledge of the flow scale description of source behavior and focuses on the large-scale properties of the count process of the aggregate traffic (number of packets in consecutive windows of size Δ).

The first part of this contribution is an experimental investigation, in the idealized loss-free situation, of Taqqu's relation linking the self-similarity of aggregate network traffic to the flow size distribution tail index. This relation had been validated only in two situations: on real internet traces, where only one point of the curve can be observed; and on simulators. Using large-scale controlled experiments generating stationary traffic traces of long duration, and using state-of-the-art estimators, I could obtain significantly better match than previously observed on simulators. I could also investigate the limits of Taqqu's model, in term of scale range involved; and clarify the role of the transport protocol in generating self-similarity. The conclusion is that in loss-free situations, Taqqu's result holds true independently of the used protocol. All of this study was performed in collaboration

with Patrice Abry, Pierre Borgnat and Guillaume Deweale, from École Normale Supérieure de Lyon (Physics lab., CNRS-ENS Lyon).

As a natural sequel of this work, I extended this experimental study to more realistic situations including losses of two types: lossy link and congestion losses. I showed that in the lossy link case, self-similarity still holds for small loss rates, but falls down for high ones. This result is not in contradiction with Taqqu's model, provided that flows are properly defined. It gives insights into what can actually be considered as a "flow" in Taqqu's model. In the congestion case, I showed that self-similarity, which cannot be measured on the total aggregate traffic saturating at the link capacity, is still present when considering the traffic of a subset of sources, although its parameter is affected by the sources' correlation introduced by their interaction at the buffer.

In the last part of this contribution, I proposed an extension of Taqqu's model, where flow rates and durations are correlated. This study was motivated by the observation of real web traffic (acquired with our tool *MetroFlux*), where such correlations introduced different tail indices for the flow size distribution and the flow duration distribution. In this case, Taqqu's model (and other existing similar ones) was unable to explain the observed self-similarity parameter. Based on a M/G/∞ like model, I introduced this correlations, and calculated the resulting self-similarity parameter. A good match was then observed with the real traffic at the origin of this development.

Estimation of the flow-size distribution's tail index from packet-sampled data. [5, 3]

In this contribution, I consider the opposite question, as compared to the previous one: from the observation of the aggregate traffic, how to recover the sources' characteristics at flow scale? I focused on the estimation of the tail index parameter of the flow size distribution, assuming a zeta distribution (a particular discrete heavy-tailed distribution). The aggregate traffic is observed at packet level. However, due to very high speeds inherent to current realistic networks, only a sub-sampled of the aggregate traffic is observed, to limit computational and storage costs.

While several solutions based on approximations existed to estimate the flow size distribution tail index α_{SI}, from sampled data, I proposed a rigorous maximum-likelihood solution. I then showed that other proposed estimators can be reinterpreted as approximations of this maximum-likelihood estimator. Using numerical simulations to assess its performance, I showed that the maximum-likelihood estimator outperforms other estimators and is notably the only one reliable at sampling rates as small as 1/1000.

This work have been done in collaboration with Florence Forbes and Stéphane Girard, from INRIA Grenoble (Mistis team).

Study of the impact of heavy-tailed flow size distributions on Quality of Service.

This contribution aims at giving answers to the question of the impact of traffic characteristics on the QoS, for both the UDP and TCP protocols. I assume a heavy-tailed distribution of the file size distribution, with some tail parameter α_{SI} (*i.e.*, the flow scale description of the source behavior is given), and I focuse on the impact of α_{SI} on the QoS, in the realistic case of finite size buffers.

I first handle the UDP case with simple matlab simulations. Concentrating on the

mean buffer occupancy and loss rate, I show that these two QoS metrics increase (are degraded) when α_{SI} decreases (with very heavy tails) for large buffers; whereas they are insensitive to α_{SI} for small buffers. I also show that the critical buffer size separating those two regimes is not the same for the mean buffer occupancy and the loss rate, and interpret this result by the different statistical nature of these two quantities.

I then turn to the TCP case, that I treat with experiments performed on our metrology platform. My results show that the impact of the α_{SI} parameter is not as clear as for the UDP case, and that more than the α_{SI} value only, the entire distribution have to be considered to evaluate buffer QoS metrics such as loss rate, delay and throughput; and end-to-end QoS metrics such as flow rate or completion time.

Identification of new multifractal scaling laws in TCP traffic and application to TCP throughput prediction.

This last contribution gives answers to the question of TCP traffic characterization at packet scale and its application to TCP throughput prediction. I show that TCP traffic intrinsically possesses a new type of scaling laws exhibiting a multifractal behavior. I show that a Markov model of the TCP congestion window is able to reproduce this new scaling behavior, and gives an analytic formula of the corresponding large deviation spectrum. The proposed model is versatile, and allows to account for various experimental conditions as well as to account for synchronization between several competing TCP flows.

This contribution is divided in two parts. The first part is a theoretical contribution. I prove an original (almost sure) large deviation theorem, which allows for the evaluation of the large deviation spectrum observed on one single realization of a stochastic process (verifying some properties). The quantity that is proven to be multifractal is not the classical Hölder exponent, but rather the mean of the process in consecutive time windows.

The second part of this contribution consists in applying this theorem to the analysis of the evolution of the congestion window of long-lived TCP sources. Using a Markovian model for TCP congestion windows, I show that TCP traffic intrinsically possesses this new kind of multifracatal scaling laws (note that this Markov model does not contradict the long-range dependence of aggregate ON/OFF traffic, because we use it only for the individual congestion window of TCP flows, and not for the aggregate traffic). This new scaling behavior is genuinely linked to the AIMD mechanism of TCP, and permits a multi-scale analysis of various TCP performance metrics, such as throughput or fairness. These properties of the traffic are experimentally validated using a large number of experiments performed on our metrology platform. I also experimentally show the capabilities of the model to account for various experimental conditions (various cross traffic), and its ability to account simply and efficiently for the synchronization between several competing TCP flows. The scaling laws that I exhibit in this contribution, and the Markov model that I proposed, permit the computation of probabilistic bounds on TCP throughput and fairness in various situations.

This work has been done in close collaboration with Julien Barral, from INRIA Rocquencourt (SISYPHE team).

1.4 Organization of the thesis

The manuscript is organized as follows.

Chapter 2 exposes the theoretical background and tools that will be used throughout the thesis. This chapter is intended to sketch the concepts used in the following, without entering into mathematical technicalities and with a network-research orientation.

In Chapter 3, we present the state of the art in network traffic and QoS analysis. This chapter exposes the different answers existing in the literature to the questions related to network traffic and QoS characterization, and points the questions for which no conclusive answer has been proposed yet, and that will be developed in the rest of the thesis.

Chapter 4 presents our large-scale experimental facility, and the tools we developed to deploy a metrology platform able to provide an experimental support to investigate traffic characteristics and QoS in realistic conditions.

In Chapter 5, we tackle the question of characterizing statistical properties of aggregate network traffic at large time scales. We start with an experimental validation of Taqqu's relation between long-range dependence and heavy tails in the idealized loss-free conditions, both with TCP and UDP. Then we move to more realistic conditions with two kinds of losses: lossy link and congestion losses. Finally, we consider the case (observed on real web traffic) where flow rates and durations are correlated.

Chapter 6 considers the opposite question: from the observation of the aggregate traffic at packet level, how to recover the sources' characteristics at flow level? To cope with very high speed constraints inherent to current realistic networks, only a sub-sample of the aggregate traffic is observed, and we then concentrate on the estimation of the flow size distribution's tail index α_{SI}.

In Chapter 7, we tackle questions related to the impact of heavy-tailed flow size distributions on QoS metrics such as buffer delay, loss and throughput; and transfer completion times or average rates.

In Chapter 8, we consider the characterization of TCP source traffic at packet level. This study is done for long-lived TCP sources, and we focus on scaling laws and their relations with the TCP mechanisms, in particular with the AIMD mechanism.

We conclude in Chapter 9, and give several perspectives for future research work.

Chapter 2

Theoretical background and tools

This chapter presents the mathematical concepts that will be used throughout the rest of the manuscript. It aims at introducing these notions rather informally through their link to network traffic. We do not enter into technical details, and give specific references for each notion.

2.1 Simple processes with small-range correlations

2.1.1 Renewal processes and Poisson process

As we have mentioned in the previous chapter, the most natural framework to describe network traffic is the theory of *point processes*, due to the inherently discrete nature of packet traffic. A point process is a mathematical construct that represents a discrete series of events as random points in space. Despite its seeming simplicity, the theory of point processes is a very rich and complex field of study in mathematics. More or less formal approaches to point processes can be found in [Cox and Isham, 1980, Daley and Vere-Jones, 2003, Baccelli and Brémaud, 2003, Lowen and Teich, 2005].

For the case of a one-dimension point process, the points are on a line corresponding to the time, and their position is often called *arrival time* and denoted by $(t_k)_{k\geq 0}$. The corresponding *counting process* N_t is the number of points between times 0 and t: $N_t = \sum_{k=0}^{\infty} \mathbb{1}_{t_k \in [0,t)}$, whereas the *count process* is the number of points in consecutive time windows of size Δ: $X_i^{(\Delta)} = \sum_{k=0}^{\infty} \mathbb{1}_{t_k \in [i\Delta,(i+1)\Delta)}$. The count process corresponds to the *increments* of the counting process: $X_i^{(\Delta)} = N_{(i+1)\Delta} - N_{i\Delta}$. Finally, the inter-arrival times are usually denoted by $\tau_k = t_{k+1} - t_k$. The simplest point process is the Poisson process.

Definition 2.1.1 (Poisson process)**.** A Poisson process of rate λ is a point process whose inter-arrival times are independent and follow an exponential distribution of parameter λ:

$$\forall k, \mathbb{P}[\tau_k \geq x] = e^{-\lambda x}, x > 0. \tag{2.1}$$

Although the term "Poisson process" usually refers to the counting process N_t (a continuous time process), we use it to denote the point process as a whole. The name "Poisson process" refers to the fact that the counting process follows a Poisson distribution of parameter λt:

$$\mathbb{P}(N_t = n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, \tag{2.2}$$

whose mean and variance are both equal to λt. The count process also follows a Poisson distribution of parameter $\lambda\Delta$. The other important property of this process is that it is independent. In other words, the counting process has independent increments. Note that these properties of the count and counting processes can also be used to define the Poisson process, and the definition 2.1.1 that we chose is then seen as a property.

The simplicity of the Poisson process comes from two assumptions: independence of the inter-arrival times and exponential distribution (the only memoryless distribution). It is the most simple example of a more general class of point processes called *renewal processes*.

Definition 2.1.2 (Renewal processes)**.** A renewal process is a point process whose inter-arrival times are independent.

Up to now, we have considered the count and counting processes where the number of points falling in some interval are counted. If each point is affected a weight, the process obtained by summing the weights of the points falling in some interval in called *renewal-reward process*. The sequence of weights is assumed independent and identically distributed (i.i.d.). Such processes can be used for example to take into account variable packet sizes, and the process $(X_i^{(\Delta)})_{i\geq 0}$ represents the bandwidth in consecutive time windows of size Δ. As compared to the Poisson process, removing the assumption of exponentially distributed inter-arrival times in the definition of renewal processes already leads to severe complications. For example, the superimposition of several renewal processes is not in general a renewal process (whereas the superimposition of Poisson processes is a Poisson process whose rate is the sum of the individual rates). The count process is also no longer independent (we shall for example see in next chapter that if the inter-arrival times are heavy-tailed distributed, it exhibits long-range dependence). The next proposition shows, however, that the superimposition of a large number of independent renewal processes converges towards a Poisson process, provided that a proper rescaling is applied (see a more general version of this proposition in [Daley and Vere-Jones, 2007, Section 11.2, p. 150]).

Proposition 2.1.3 (Convergence of the superimposition of independent renewal processes towards a Poisson process)**.** *Let N be a renewal process on $\mathbb{R}$, whose inter-arrival-time distribution has finite mean λ^{-1}. Let N_n be the point process obtained by superimposing n independent replicates of N and dilating the scale of $\mathbb{R}$ by a factor n. Then as n goes to infinity, N_n converges to a Poisson process of rate λ.*

2.1.2 Markov chains

In the previous section, we considered the point process corresponding to packet arrivals. Such a packet-scale description is not always necessary, and many results can be obtained using a flow-scale description: the point process now corresponds to the flow arrivals, and each flow is assumed to emit continuously at a constant rate during a random time. This situation corresponds to the classical setting used in queueing theory, where the flows are usually called *jobs*, and the flow duration called *service time*. It also represents the case of phone networks, where the flows are the calls.

In the case where the job- (flow-) arrival process is Poisson, and the service times (flow durations) are exponentially distributed (the case of an M/M/1 queue), the evolution of

the number of customers X_k (number of active flows) at time $t = k\Delta$ can be described by a Markov chain.

Markov chains is now a rather classical topic in mathematics, and a lot of references are available on the subject. The reader can consult for example [Feller, 1971, Revuz, 1984, Benaïm and El Karoui, 2005], and the references therein for more information.

Definition 2.1.4 (Markov chain)**.** A discrete process $(X_k)_{k\geq 0}$ is a Markov chain if

$$\mathbb{P}(X_{k+1} = n_{k+1}|X_k = n_k, X_{k-1} = n_{k-1}, \cdots, X_0 = n_0) = \mathbb{P}(X_{k+1} = n_{k+1}|X_k = n_k). \quad (2.3)$$

This is the simplest case of correlated process: the probability of an event depends only on the preceding event, and not on an event occurred further in the past. In other words, one often says that "the future depends only on the present". As for the case of point processes, Markov chains reveal many rich behaviors despite their apparent simplicity.

The set of possible values for X_k is called the *state space* E. If we assume that there is a maximal number of active flows in the queue (for example because of the finite number of sources), then process X can only take a finite number of values and the state space is finite. We present only this case here and assume that the state space is $\{1, \cdots, c\}, c \geq 1$. The Markov chain is characterized by a *transition matrix* Q, such that:

$$Q_{ij} = \mathbb{P}(X_{k+1} = j|X_k = i). \quad (2.4)$$

The properties of the Markov chain are governed by those of the matrix Q. This matrix always verifies $\forall i, \sum_{j=1}^{c} Q_{ij} = 1$. Given the distribution $p^{(k)}$ on E at time k, the transition matrix determines the distribution at time $(k+1)$: $p_j^{(k+1)} = \sum_{i=1}^{c} p_i^{(k)} Q_{ij}$, usually written in short $p^{(k+1)} = p^{(k)}Q$ and named *Chapman-Kolmogorov equation.* A steady-state or invariant distribution is a fixed point of this equation:

Definition 2.1.5 (Steady-state distribution)**.** A distribution p on E is called steady-state distribution (or invariant distribution) of the Markov chain X of transition matrix Q if

$$p = pQ, \ \textit{i.e.} \ \forall j \in E, p_j = \sum_{i=1}^{c} p_i Q_{ij}. \quad (2.5)$$

A steady-state distribution always exists, but the irreducibility assumption is necessary for its unicity:

Definition 2.1.6 (Irreducible Markov chain)**.** A Markov chain X of transition matrix Q is called irreducible if $\forall (i,j) \in E^2$, there exists $k \geq 1$ such that $(Q^k)_{ij} > 0$.

A Markov chain is irreducible if each state of E is accessible from any state of E. It means that the chain cannot be reduced, or separated in two independent chains. If the chain is irreducible, there exists a unique steady state distribution, verifying $\forall i \in E, p_i > 0$. An almost-sure ergodic theorem also holds:

Theorem 2.1.7 (Almost-sure ergodic theorem)**.** *If X is an irreducible Markov chain of steady-state distribution p, then almost surely:*

$$\lim_{n\to} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_{\{X_k = i\}} = p_i. \quad (2.6)$$

This theorem means that when one realization of the Markov chain is generated, the result obtained by counting the proportion of a particular value converges toward the steady-state distribution of this value as the size of the realization grows. It implies that the ensemble mean of the process is equal to the empirical mean calculated on one realization. When the chain is aperiodic (Definition 2.1.8), a stronger ergodic theorem can be stated, implying another type of convergence.

Definition 2.1.8 (Aperiodic Markov chain)**.** Let $i \in E$ and $R(i) = \{k \in \mathbb{N} : Q^k(i,i) > 0\}$. The period of i is the greatest common divider of $R(i)$. If Q is irreducible, all the states have the same period, and the Markov chain is called aperiodic if this period is one.

The theory of Markov chains has been extensively used to dimension and control phone networks, following the work of Erlang in 1917. At the beginning of packet networks, it was also used [Kleinrock, 1969] to model the traffic. However, the assumption of exponentially distributed flow durations (underlying the direct use of simple Markov chain models) was proven to fall down in the Internet in the early 90s. Instead the flow-duration distributions appeared to exhibit heavy-tailed distributions. While modeling the aggregate traffic with Markov processes in such a situation is still theoretically possible, the state space becomes extremely large and so complicated that more parsimonious models are usually preferred to account for the long-term correlations. Such models, based on long-range dependence are presented in Section 2.3. To model the traffic of an individual source, however, Markov chain models can be accurately used with very simple state spaces, representing the possible values of the congestion window. Many such models have appeared since the late 90s. The idea is that with Reno's AIMD mechanism, the new congestion window depends only on the current value: it is either divided by two or increased by one, with some probabilities related to the loss probability.

2.2 Heavy-tailed distributions

There are various ways to define classes of distributions decreasing more slowly than the exponential distribution [Adler et al., 1998, Markovich, 2007]. Heavy-tailed distributions are one such class and we choose here a classical definition adapted to our study, which is sometimes named "regularly varying tail distributions" in the specialized literature.

Definition 2.2.1 (Heavy-tailed random variable)**.** A random variable Z is *heavy-tailed* with tail index α if its distribution is of the form

$$P_Z(|Z| \geq z) = z^{-\alpha} L(z), \tag{2.7}$$

where L is a slowly varying function, *i.e.* $L(tz)/L(z) \to 1$ as $z \to \infty$ for any $t > 0$.

In this thesis, we consider positive random variables only, and we remove the absolute value of equation (2.7). The term "heavy-tailed" can apply either to the random variable or to its distribution. With this definition, not all the distributions decreasing more slowly than the exponential correspond to heavy-tailed random variables. For example, the Weibull distribution is not heavy-tailed. As compared to the exponential distribution of equation (2.1) governed by the characteristic value λ, there is no such characteristic value in the heavy-tail distribution of equation (2.7).

The most popular paradigm of heavy-tailed distribution is the *Pareto distribution*. While a "pure Pareto" distribution is proportional to $z^{-\alpha}$, we will use throughout the manuscript a distribution of the form:

$$P_Z(Z \geq z) = k^\alpha (z+k)^{-\alpha}, z > 0, \tag{2.8}$$

where k is a positive parameter, and $\alpha > 0$. The corresponding normalized pdf is

$$p_Z(z) = \frac{\alpha k^\alpha}{(z+k)^{(\alpha+1)}}. \tag{2.9}$$

This distribution has two parameters: k and α. The parameter k is a "location" parameter: changing the value of this parameter roughly translates the distribution on the x-axis. For a given value of α, this parameter controls the mean of the distribution, which reads $\mathbb{E}Z = \frac{k}{\alpha-1}$ (if it is finite). Parameter α controls the maximal order moment which is finite:

$$\alpha = \sup_r \{r > 0 : \mathbb{E}(Z^r) < \infty\}, \tag{2.10}$$

If $\alpha < 1$, the mean is infinite, if $1 < \alpha < 2$, the mean is finite but the variance is infinite, and if $\alpha > 2$ the variance is finite. This parameter is also the exponent of the algebraic decrease of the distribution, in the asymptotic regime of large values. Because it will be useful for the estimation of α, we mention here that by a duality argument [Gonçalves and Riedi, 2005], the power-law decay of the distribution (2.8) for large values transposes to a power-law decay of the characteristic function $\chi_Z(s) = \mathbb{E}e^{-isZ}$ at the origin; and the tail exponent α correspondingly transposes according to:

$$\alpha = \sup_r \{r > 0 : 1 - \Re\chi_Z(s) = \mathcal{O}(s^r) \text{ as } s \to 0^+\}, \tag{2.11}$$

where $\Re$ stands for the real part. This power-law decay of $\Re\chi_Z(s)$ reflects a scale-invarient property, a concept that we now expose.

2.3 Scaling laws

Scaling laws can be related to several concepts such as self-similarity or multifractality. Even though these notions are not equivalent, they share a common property: there is no characteristic scale that prevails to describe the system. Instead, different scales of the system are linked together by a specific scale-invariant property. Scaling laws are also closely related to the concepts of fractal dimensions, but we do not use this angle in our presentation.

For more than half a century now, scaling laws have been the object of an extremely large number of research activities, from both the theory and application sides. Our short presentation is then necessarily oversimplified and far from exhaustive. We focus only on the concepts used in the rest of the manuscript. The interested reader can consult the collection of chapters covering a large number of topics in [Abry et al., 2002b], and more specialized references on mathematical aspects and on applications [Beran, 1994, Mandelbrot, 1997, Park and Willinger, 2000, Embrechts and Maejima, 2002, Doukhan et al., 2003].

2.3.1 Self-similarity and long-range dependence

Self-similarity and fractional Brownian motion

Self-similarity is the simplest example of scaling laws, expressing that a process remains statistically the same when rescaled by any factor c, provided that it is re-normalized by c^H:

Definition 2.3.1 (Self-similarity)**.** A process $(Y(t))_{t\in\mathbb{R}}$ is self-similar of index H is

$$\{Y(t), t \in \mathbb{R}\} \overset{f.d.d.}{=} \{c^H Y(t/c), t \in \mathbb{R}\}, \forall c > 0, \tag{2.12}$$

where $\overset{f.d.d.}{=}$ stands for finite-dimension distribution equality.

Scale-invariant quantity H, called Hurst parameter, characterizes the link between different scales of the process. This definition implies that for all finite moments of Y: $\mathbb{E}|Y(t)|^q = \mathbb{E}|Y(1)|^q|t|^{qH}$, clearly exhibiting the two major characteristics of a self-similar process: its scale-free nature and its necessary non-stationarity.

A sub-class of the class of self-similar processes is that of stationary increments processes:

Definition 2.3.2 (Stationary increments)**.** A process Y has stationary increments if $\forall\Delta \in \mathbb{R}$, the law of the increment process:

$$X^{(\Delta)} = \{X^{(\Delta)} = Y(t + \Delta) - Y(t), t \in R\}, \tag{2.13}$$

is independent of t, *i.e.*

$$Y(t + \Delta) - Y(t) \overset{f.d.d.}{=} Y(\Delta) - Y(0). \tag{2.14}$$

Self-similar processes of index H whose increments are stationary are usually written in short H-sssi. Such processes necessarily have an index H in $(0, 1)$ and an auto-covariance structure of the form (if the second order moment is finite):

$$\mathbb{E}Y(t)Y(s) = \frac{\mathbb{E}Y(1)^2}{2}(|t|^{2H} + |s|^{2H} - |t - s|^{2H}),\ \forall(t, s) \in \mathbb{R}. \tag{2.15}$$

A corresponding equation exists for the auto-covariance function of the increment process.

The most prominent example of H-sssi process is the fractional Brownian motion [Mandelbrot and Van Ness, 1968]:

Definition 2.3.3 (Fractional Brownian motion)**.** The fractional Brownian motion of index H is the only H-sssi processes which is Gaussian and centered.

In this definition, unicity arises from the fact that a Gaussian process is entirely determined by the auto-covariance structure, here of the form of equation (2.15). If $H = 1/2$, the fBm is then an ordinary Brownian motion. The increment process of the fBm is called fractional Gaussian noise (fGn). If $H > 1/2$, it is a particular case of process possessing the less restrictive long-range-dependence property. Before defining long-range dependence, let us mention that the Gaussian property is not necessary for the self-similar property of Definition (2.3.1) to hold. Another important class of self-similar processes than fBm, is that of α-stable Lévy motions. These Lévy motions are defined as processes with stationary independent increments of α-stable distribution ($0<\alpha<2$) [Samorodnitsky and Taqqu, 1994]. Such processes can be shown to be H-sssi, with $H = \frac{1}{\alpha}$. Note however that in the case of infinite variance processes, relation (2.15) does not hold.

long-range dependence

Definition 2.3.4 (Long-range dependence)**.** A second-order stationary process X is long-range dependent if its correlation function statisfies:

$$\mathbb{E}X(t)X(t+\tau) \underset{\tau\to\infty}{\sim} C|\tau|^{-\beta}, \text{ for } 0<\beta<1, \tag{2.16}$$

where C is a constant.

The long-range dependence is then a second-order asymptotic (for large scales) property. The fGn is long-range dependent with parameter $\beta = 2-2H$, if $H > 1/2$. More generally, for any H-sssi process of finite variance such that $H > 1/2$, the increment process $X^{(\Delta)}$ is long-range dependent. Strictly speaking, long-range dependence and self-similarity designate two different notions, although closely related: the latter is associated to non-stationary time series, such as fBms, while the former is related to stationary time series, such as fGn. Moreover, whereas self-similarity implies long-range dependence, the opposite does not hold (long-range dependence is only equivalent to asymptotic second order self-similarity). As for the case of self-similarity, the Gaussian property is not necessary for the long-range dependence property of equation (2.16) to hold, and any distribution is acceptable as long as its second order moment is finite, so that the auto-covariance function is well defined. For infinite variance processes that are H-sssi, one cannot deduce a long-range dependence property of the increments (Lévy motion for example have independent increments).

Equation (2.16) clearly exhibits the slow decrease of the auto-covariance function, which is not integrable. This long-term correlation structure is in sharp contrast with the short-term correlations of Markov chains.

2.3.2 Small scales: Hölder exponent and multifractality

The fBm/fGn have been massively used to model long-range-dependence effects in network traffic. However, its uniform regularity at small scales limits its application in several situations. It has led to the introduction of a new class of scaling laws that we describe now: Hölder multifractality.

Hölder exponent and local regularity

Studied process Y is now defined on a bounded interval, that without loss of generality we assume to be $[0,1]$. At each scale $n \in \mathbb{N}$, the interval $[0,1]$ is divided in 2^n consecutive basic intervals of size 2^{-n}. Exploring the small-scales properties of the signal then corresponds to letting n go to infinity.

Definition 2.3.5 (Hölder exponent)**.** At scale n, the *grain Hölder exponents* are defined for $j = 0, \cdots, 2^n - 1$ by:

$$h_n(j) = \frac{\log \sup\{|Y(u)-Y(v)| \ : \ u,v \in [j2^{-n},(j+1)2^{-n})\}}{\log 2^{-n}}, j = 1, \cdots, k_n. \tag{2.17}$$

The *Hölder exponent* at time t is defined by

$$h(t) = \liminf_{n\to\infty} h_n(\lfloor 2^n t \rfloor) \tag{2.18}$$

This definition of the Hölder exponent based on the maximal *oscillation* of the process within an interval of size 2^{-n} coincides with the classical definition based on the degree of local regularity in most cases. The Hölder exponent characterizes the local regularity of the process at time t: values close to 0 denote almost discontinuous processes, whereas values close to 1 are associated to almost differentiable processes. Other local exponents exist, in particular based on the wavelet coefficients of the signal, that we do not consider here (see [Riedi, 2003]).

For a fBm of Hurst parameter H, the Hölder exponent is constant: $h(t) = H$ almost surely. The fBm is then called *monofractal*. However, for more complex processes called *multifractal*, the Hölder exponent erratically varies. There exists different possible definitions of the singularity spectrum which characterize the variability of the Hölder exponent. In many cases, the "multifractal formalism" is said to hold, meaning that all spectra coincide. We describe here only the large-deviation spectrum.

large-deviation spectrum

Definition 2.3.6 (large-deviation spectrum)**.** The large-deviation spectrum is defined as

$$f(\alpha) = \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log_2 \left[\#\left\{ j \in \{0, \cdots, 2^n - 1\} : \; h_n(j) \in [\alpha - \varepsilon, \alpha + \varepsilon) \right\} \right]. \tag{2.19}$$

Note that the existence of the limit on n is not trivial (see [Riedi, 2003] for conditions guaranteeing its convergence, and see also similar arguments developed in Chapter 8). With this definition, the large-deviation spectrum is $-\infty$ for values of the Hölder exponent never achieved, or between 0 and 1 in the domain of observable Hölder exponents. Equation (2.19) describes the scale invariance of the statistical repartition of grain Hölder exponents and permits an easy interpretation: roughly, when n goes to infinity, the proportion of grain Hölder exponent values around α decreases as $\exp(n(f(\alpha) - 1))$, where $f(\alpha)$ does not depend on n.

2.4 Wavelets and estimation

In previous sections, we introduced essential quantities to describe network traffic: tail index α (usually of the flow-size distribution), Hurst parameter H, and the large-deviation spectrum (usually of the aggregate traffic). We now present techniques to estimate these parameters.

2.4.1 Discrete wavelet transform

We first briefly describe wavelets, because they constitute a convenient tool to identify scale-invariance properties, and they are then used for estimation issues. More information on wavelets and their numerous applications can be found in [Meyer, 1992, Daubechies, 1992, Mallat, 1999].

Performing a wavelet transforms of a signal X consists in computing the (Discrete) Wavelet Transform coefficients:

$$d_X(j,k) = \langle \psi_{j,k}, X \rangle, \tag{2.20}$$

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 $L^2(\mathbb{R})$. Reference template ψ_0 is termed mother-wavelet and is characterized by its number of vanishing moments $N_\psi > 1$, an integer such that $\int t^k \psi_0(t)\, dt = 0, \forall k = 0, ..., N_\psi - 1$. The wavelet coefficient is the inner product of signal X with a translated and dilated version of the mother-wavelet. Integer j is usually called the *scale* and corresponds to the dilatation of the analysis wavelet at scale j. This dilation mechanism intuitively justifies the appropriateness of the wavelet transform in the study of many phenomena implying scale-invariance properties.

2.4.2 Estimation of the tail index

Estimation of the tail exponent α for heavy-tailed random variables is an intricate issue that received considerable theoretical attention in the statistics literature [Seal, 1952, Hill, 1975, McCulloch, 1997, Crovella and Taqqu, 1999, Nolan, 2001, Gonçalves and Riedi, 2005, Markovich, 2007, Clauset et al., 2009]: measuring the tail exponent of a heavy-tailed distribution amounts to evaluating, from observations, how fast the probability of rare events decreases in equation (2.7). We present here only the method of [Gonçalves and Riedi, 2005] which is a robust wavelet-based method, used in this thesis for the first time in the context of network traffic. An advantage of this estimator is that it is non-parametric, *i.e.*, it does not assume *a priori* a specific form of the distribution.

The principle of this estimator is simple and relies on the wavelet identification of the scale-invariance property associated with the power-law decrease of the characteristic function $\Re\chi_Z(s)$ at the origin, observed in equation (2.11). Hence, computing the discrete wavelet decomposition of $\Re\chi_Z$, and retaining only the wavelet coefficients that lie at the origin $k = 0$, yields the following multi-resolution quantity:

$$d_{\chi_Z}(j, 0) = \mathbb{E}\Psi_0\left(2^j Z\right) \leq C2^{j\alpha} \text{ for } j \to -\infty, \tag{2.21}$$

where $\Psi_0(\cdot)$ denotes the Fourier transform of analyzing wavelet $\psi_0(\cdot)$. Now, let $\{z_0, \cdots, z_{n-1}\}$ be a set of i.i.d. heavy-tailed random variables of tail index α, and replace the ensemble average in equation (2.21) by its empirical estimator, the estimate $\widehat{\alpha}$ simply results from a linear regression of the form

$$\begin{aligned} \log \widehat{d}^{(n)}_{\chi_Z}(j, 0) &= \log n^{-1} \sum_{i=o}^{n-1} \Psi(2^j z_i) \\ &\approx \widehat{\alpha} j + \log C, \text{ as } j \to -\infty. \end{aligned} \tag{2.22}$$

The estimator was proven to converge for all heavy-tailed distributions, with a reduced variance of estimation in $\mathcal{O}(n^{-1})$ (where n is the sample size). It was moreover shown to be robust for all $\alpha > 0$, even for α close to 1^+, when the apparent data mean instability is responsible for the slow convergence of usual empirical estimates. We refer the interested reader to [Gonçalves and Riedi, 2005] 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 of equation (2.22) holds, and which is helpful for practitioners to adequately adjust their linear fitting over a correct scale range.

2.4.3 Estimation of the Hurst parameter

Many methods exist to estimate the Hurst parameter from a sample of the process, such as the Variance-Time method, the R/S method, or the Whittle method (see [Beran, 1994]). However, all of them have been shown to be outperformed by the method that we present here: the wavelet method [Abry et al., 1995]. This method turned out to be particularly efficient at analyzing Internet time series in [Abry et al., 2000, Abry et al., 2002a] and has then been massively used in this context.

The method is based on a discrete wavelet decomposition of signal X (see the previous section). Then, if the process is H-sssi, the variance of the wavelet coefficients verifies:

$$\mathbb{E}|d_X(j,k)|^2 = \mathbb{E}|d_X(0,0)|^2 2^{j(2H+1)}, \tag{2.23}$$

and, provided that $N_\psi > H + 1/2$, the sequence $\{d_X(j,k),\, k = \ldots, -1, 0, 1, \ldots\}$ forms a stationary and weakly correlated time series. These two central properties warrant using 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 $\mathbb{E}|d_X(j,k)|^2$. Equation (2.23) 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). A (weighted) linear regression of the LD within a proper range of octaves j_1, j_2 is used to estimate H.

The wavelet estimator is endowed with a practical robustness that comes from its extra degree of freedom N_ψ. In the case of a Gaussian process, the wavelet estimator is also able to provide error bars on the estimated value of the Hurst parameter. In practice, the main difficulty lies in the correct choice of the regression range $j_1 \leq j \leq j_2$, which basically should be in accordance with the scale range in which the correlation structure of equation (2.16) holds. This will be discussed throughout the manuscript, in the light of actual measurements.

2.4.4 Estimation of the large-deviation spectrum

Because multifractality corresponds to a scale-invariance property, wavelet-based estimators can also be used to estimated the large-deviation spectrum. However, we did not use such a tool in our work. Instead, we use a technique based on the direct computation of the oscillations of equation (2.17). Then, from a practical viewpoint, estimation of the large-deviation spectrum is a delicate issue. The main problem is to adapt quantity ε of equation (2.19) to the scale n. In this work (estimations in Chapter 8), we use a recent original idea of [Barral and Gonçalves, 2009], that we do not further develop here.

Chapter 3

Network traffic and Quality of Service analysis: state of the art

Introduction

This chapter presents the state of the art in network traffic and QoS analysis, from both the theoretical and experimental sides. We review existing answers to the general questions of the introduction, and we point particular questions which remain open and for which we will propose answers in the rest of this manuscript.

We first present the literature on the statistical analysis of aggregate network traffic (Section 3.1). We focus here on the analysis of scaling laws, at both large and small scales. Section 3.2 presents the state of the art on the analysis of Quality of Service in performance terms, and more particularly on the impact of scaling laws and heavy-tails on QoS. Finally, we expose in Section 3.3 the literature on the analysis of TCP-traffic properties at packet scale and TCP throughput evaluation.

For the sake of clarity, the literature related to the flow size distribution's tail index estimation from sampled data has been voluntarily excluded from this state of the art, and will be thoroughly exposed in the dedicated chapter (Chapter 6).

3.1 Statistical properties of aggregate network traffic

Aggregate network traffic characterization.

Statistical characterization of aggregate network traffic has been the object of a large quantity of research activity. Many recent results can be found in books [Adler et al., 1998, Park and Willinger, 2000, Doukhan et al., 2003] and survey papers [Abry et al., 2002a, Erramilli et al., 2002]. Our presentation is focused on the analysis of scaling laws, that we divide in two categories: large-scale and small-scale properties.

3.1.1 Large scales

Observation

Discovery of the self-similar nature of aggregate traffic of various nature. A major breakthrough in network traffic analysis has been the discovery of its self-similar property at large time scale [Leland et al., 1993,Paxson and Floyd, 1994] (see also extended versions of these papers in [Leland et al., 1994, Paxson and Floyd, 1995]), invalidating

Markov models that had been successfully used for phone networks traffic. In [Leland et al., 1993, Leland et al., 1994], the authors study aggregate traffic traces acquired between 1989 and 1992 in a Local Area Network (LAN) at the Bellcore Morris Research and Engineering center. Using Variance-Time, R/S and periodogram methods, they show that the averaged bandwidth in consecutive time windows (*i.e.*, the count process) exhibits long-range dependence. In [Paxson and Floyd, 1994, Paxson and Floyd, 1995], the authors show that traffic acquired in various Wide Area Networks (WANs) between 1989 and 1991 also exhibits long-range dependence. They use the Variance-Time and Whittle methods, also applied to the averaged bandwidth. Using the same methods, the authors of [Beran et al., 1995, Garrett and Willinger, 1994] show that Variable Bit Rate (VBR) traffic also exhibit the same long-range dependence property. In all of these papers, the authors posit the presence of some heavy-tailed distribution, mainly that of the flow size, as the origin of the long-range dependence.

Taqqu's relation between self-similarity and heavy-tails. Following up this seminal discovery, the authors of [Willinger et al., 1995, Willinger et al., 1997] propose a physical explanation for the presence of long-range dependence. They consider an ON/OFF model and propose the relation:

$$H = \frac{3 - \alpha_{\mathrm{H}}}{2}, \tag{3.1}$$

where $\alpha_{\mathrm{H}} = \min(\alpha_{\mathrm{ON}}, \alpha_{\mathrm{OFF}}, 2)$, α_{ON} (respectively α_{OFF}) being the tail index of the distribution of ON (respectively OFF) periods. This relation confirms that heavy-tailed ON/OFF periods are able to generate long-range dependence in the aggregate-traffic count process, and gives an analytical formula linking both properties. It also confirms the role of infinite variance distributions in generating long-range dependence: if $1 < \alpha_{\mathrm{ON}} < 2$, then the ON period distribution has infinite variance, and $1/2 < H < 1$, *i.e.*, the traffic is long-range dependent (and similarly for the OFF periods). In the following, we will often call relation (3.1) "Taqqu's relation", in reference to the paper [Taqqu et al., 1997b] where its proof appeared.

Validation of Taqqu's relation. This relation have been validated in two different situations. The first situation is that of [Crovella and Bestavros, 1996, Crovella and Bestavros, 1997, Crovella et al., 1998], where the authors analyze World Wide Web (WWW) requests' traffic traces acquired in 1995. They find that the distribution of the observed ON times is heavy-tailed with a tail parameter around 1.21, while the OFF-time distribution is also heavy-tailed of index 1.5. Using Variance-Time, R/S and periodogram methods, they estimate a Hurst parameter around 0.8, whereas relation (3.1) predicts value 0.89, but no clear interpretation of this deviation is given. Interestingly, in these papers, the authors find that the file-size distribution has a heavier tail than the ON-size distribution (a tail parameter around 1.06). They explain this difference by slow-start effects, giving a higher rate to larger flows, and thus decreasing the tail of the ON-times distribution. No comment, however, is made on the fact that relation (3.1) assumes constant rates and might not be valid when the tail parameters of the file sizes and ON-times distributions have different values. Finally, the authors claim that the heavy-tailed distribution of the files transfered in their traces simply results from the heavy-tailed distribution of the available files in the web.

Another type of validation of relation (3.1) is presented in [Park et al., 1996]. The authors use simulator ns2 to generate ON/OFF-type traffic with different tail indices, and estimate the corresponding Hurst parameter using Variance-Time and R/S methods. Using heavy-tailed file sizes and exponential OFF times with the TCP protocol, they find that the Hurst parameter decreases when the tail parameter increases, but a quite large deviation is observed with respect to relation (3.1). In the reverse case (when the file sizes are exponential and the OFF times heavy-tailed), they obtain a Hurst parameter no larger than 0.7 with a tail parameter of 1.05. In these experiments, loss rates are not controlled and vary with the tail parameters, which might have an impact on the scaling properties of the traffic (a quite high loss rate of 4% is observed in the worst case where the tail index is 1.05). Similarly, when a greedy UDP protocol is used, the Hurst parameter is no larger than 0.7. They interpret this last observation by the fact that a greedy protocol concentrates the transfer mass in time and reduces the dependence structure at large scales, but no bound is given to explain at which scales the long-range dependence can be observed. Apart from the partially controlled environment, another origin of the mismatch observed in [Park et al., 1996, Crovella and Bestavros, 1996] might lie in the statistical tools used for the estimation of the Hurst parameter. Indeed, Variance-Time and R/S methods have been shown to suffer from many drawbacks, and the wavelet estimator proposed in [Abry and Veitch, 1998] has been proven to be a much more robust tool. However, even with this robust estimator, validation of relation (3.1) is not straightforward (see [Roughan and Veitch, 2007] and reference therein), and the scale range for estimating the Hurst parameter has to be chosen carefully. Only very recently, relation (3.1) has received a validation where the Hurst parameter is estimated with the wavelet estimator, carefully used [Abry et al., 2009]. In this work, the authors use M/G/∞ numerical simulations (with matlab), and ON/OFF simulations (with ns2) and focus on the signal processing aspect. They show that the regression scale range used in the wavelet estimation has to be chosen carefully and give plausible explanation for the deviation commonly observed in the validation of relation (3.1).

The role of TCP. Despite the meaningful contribution of [Willinger et al., 1995, Willinger et al., 1997] and the physical explanation that it offers to explain large-scale self-similarity in aggregate network traffic, another possible origin of this property has been debated over the last decade: the TCP protocol. In [Veres and Boda, 2000, Sikdar and Vastola, 2001], based on ns2 simulations of long-lived TCP connections, the authors argue TCP control mechanisms are able to generate long-range dependence in the aggregate traffic, independently of the ON/OFF structure of the sources. However, the scaling behavior is not observed at large time scales and corresponds to *pseudo self-similarity*, *i.e.*, non-asymptotic self-similarity lying in some intermediate scale range. A plausible interpretation of this pseudo self-similarity observed in long-lived TCP connections is given in [Guo et al., 2001, Guo et al., 2002, Figueiredo et al., 2002]. The authors use a Markov model to show that under very high loss rates (around 20%), the timeout mechanism can generate heavy-tailed distributed OFF periods. The limited support of this OFF-period distribution implies that resulting self-similarity is observable on a limited scale range only and thus justifies its non-asymptotic character. Then, the same simulations as in [Veres and Boda, 2000] are performed in [Figueiredo et al., 2005] over longer durations, and show that the aggregate traffic is asymptotically not long-range dependent at large time scales.

This way, they point again a classical pitfall in estimating long-range dependence: a bad choice of scale range. Based on ns2 simulations, the authors of [MacIntyre et al., 2008] show that this kind of pseudo scaling behavior holds independently of the TCP variant used. To conclude on the link between TCP and self similarity, let us finally mention paper [Veres et al., 2000]. In this paper, the authors do not debate about how TCP can generate self-similarity, but they show that TCP can propagate self-similarity. They consider the case where TCP traffic shares a fixed-capacity link with self-similar cross traffic. Due to the control mechanisms, the sum of the TCP traffic and the cross traffic saturates at the link capacity and the TCP traffic therefore inherits the self-similar property of the cross traffic.

Questioning the evidence for self-similarity in backbone traffic. At the approximate same time as this debate on the role of TCP in generating self-similarity, a few papers appeared questioning the evidence for self-similarity in Internet backbones. In [Cao et al., 2001b, Cao et al., 2001a], the authors use traffic traces from 6 different Internet backbone monitors measuring 15 interfaces ranging from 100 Mbps to 622 Mbps, and show that this traffic is well-modeled by a Poisson process (thus not long-range dependent). They argue that the high statistical multiplexing on backbone links is responsible for the Poisson behavior. This apparent contradiction is solved somehow in the same way as for the role of TCP in generating self-similarity: this is a matter of time scales. In [Roughan and Gottlieb, 2002], the authors indeed show that backbone traffic is long-range dependent if investigated at sufficiently large scales (the traces used in [Cao et al., 2001b, Cao et al., 2001a] are at most 5 minutes long, whereas in [Roughan and Gottlieb, 2002], one-year-long traces are used). Similar findings are reported in [Karagiannis et al., 2004a, Karagiannis et al., 2004b], and finally, a very recent comprehensive study of seven years' worth of traces (acquired on a trans-Pacific backbone link: the MAWI dataset) confirms that the long-range dependence property is present even for highly multiplexed traffic [Borgnat et al., 2009].

The literature related above shows that the conditions underlying the observation of self-similarity in aggregate traffic, and more particularly the validity of Taqqu's relation (equation 3.1), have been largely debated in the last decade. However, no comprehensive study performed on a real large-scale experimental facility have yet been realized, and some points remain unclear:

- To what extend is Taqqu's relation valid in real network conditions? What are the scales involved? Does the protocol play a role? In which situation (loss-free, lossy link, congestion)?

Self-similarity of flow arrivals. All the previously mentioned papers where focusing on the long-range dependence of the averaged bandwidth of the packet traffic (*i.e.*, the count process) and its origin, and one might now wonder whether the flow-arrival process possesses similar properties. This question is tackled in [Hohn et al., 2002b], where the authors show, based on various publicly available traces, that the count process corresponding to the flow-arrival process (*i.e.* the number of flows in consecutive time windows) indeed exhibits self-similarity. The same authors conclude however in subsequent papers [Hohn et al., 2002a, Hohn et al., 2003] that this long-range dependence in the flow-arrival process is not responsible for the long-range dependence in the packet-arrival process. The method

used in these papers consists in performing what they call *semi-experiments*: based on a real Internet trace, some characteristics are artificially modified, while keeping others constant. For example, the authors modify the flow-arrival process to make it Poisson, while keeping the same packet-arrival processes within each flow. They observe the same large-scale scaling behavior and conclude that the correlation structure of the flow-arrival process is not responsible for the observed long-range dependence on the packet-level traffic. Such trace-driven analysis, while essential to understand key properties of Internet traffic, often leads to partial results intimately linked to the specific characteristic of the studied trace, that might fall down when considering other traces with different characteristics. To understand, in more generality, the different mechanisms susceptible to generate self-similarity , we now turn to the different models that have been proposed in the literature.

Models

A large number of models able to reproduce large-scale self-similarity have been proposed and applied in various fields of science [Doukhan et al., 2003]. We present here only those which have been applied to network-traffic analysis. We give a network-oriented presentation of these models and their links, and we do not enter mathematical details of their proofs and precise statement.

Basically, all of these models rely on the introduction of a heavy-tailed distribution. We can distinguish three main categories of models whose parameters are closely related to meaningful network traffic parameters. The first model is the *fractal point processes* (FPP) model. It considers renewal point processes whose inter-arrival times are heavy-tailed. Even though it is not the most used in network-traffic analysis, we present it first because it is the simplest and it is useful to understand certain properties of the other models. The second model considers renewal reward processes, whose inter-arrival times are also heavy-tailed. The ON/OFF model is a special case of this model, in which the reward strictly alternate between 1 and 0. The third class of models is the infinite source Poisson models with heavy-tailed transmission times. This class of models include $M/G/\infty$ models, Poisson shot-noise models and cluster point processes (with Poisson cluster arrivals and heavy-tailed cluster sizes).

Fractal point processes. Fractal point process (FPP) are point processes whose count process exhibit long-range dependence. There are many kinds of such processes (see [Ryu and Lowen, 1996, Lowen and Teich, 2005]), but we introduce here only the simplest and most closely related to network traffic: the fractal renewal process (FRP). The FRP has been introduced in [Lowen and Teich, 1993]. It is a simple renewal process, whose inter-arrivals are heavy-tailed with tail index α. If $1 < \alpha < 2$, then the count process is long-range dependent, and its Hurst parameter satisfies equation (3.1), with $\alpha_{\mathrm{H}} = \alpha$. In [Lowen and Teich, 1993], the authors consider a pure Pareto distribution with a lower cutoff A (*i.e.* proportional to $1/x^{\alpha}$ for $x > A$), and show that the algebraic decrease of the autocorrelation function is observed at time scales much larger than A. This result allows us to make one important remark: generally speaking, the correlation structure of the inter-arrival process is not the same as the correlation structure of the count process. Indeed, here the inter-arrivals are independent (this is a renewal process), whereas the count process exhibits long-range correlations. However, the sharp difference between the correlation structures of both processes observed here is likely due to the heavy-tailed

distributed inter-arrival times. Indeed, if inter-arrivals are exponentially distributed and the count process exhibits long-range dependence, the process cannot be a renewal process (otherwise it would be a Poisson process) and the correlation structure of the count process must then be somehow present in the inter-arrival process. This remark is important when considering the superimposition of a large number of independent FRPs, whose inter-arrival distribution is maintained the same (with the same lower cutoff A). The resulting process is no longer a renewal process. Intuitively, the inter-arrival time distribution tends toward an exponential distribution (see also [Lowen and Teich, 2005]). However the correlation structure of the count process is unchanged (it remains long-range dependent) at time scales much larger than A, even if the number of points at such scales increase with the number of superimposed processes. Despite its exponentially distributed inter-arrivals, the superimposed process should not be confused with a Poisson process. This remark will be important in understanding the difference between the renewal models and the M/G/∞ models. Before exposing these models, let us mention that applications of the FRP model to network traffic (the point process then corresponds to the packet-arrival process) are presented in [Ryu and Lowen, 1996], and in [Paxson and Floyd, 1995].

Renewal models. Renewal models are based on the same general setting firstly introduces by Mandelbrot in 1969 [Mandelbrot, 1969] in an economical context. They consider a renewal reward process, whose inter-renewal times are denoted by U_k and rewards are denoted by W_k. The sequences $\{U_k\}$ and $\{W_k\}$ are always assumed i.i.d. and mutually independent. The process of interest $W(t)$ is then set equal to W_k if t lies in the kth renewal period. In a network context, one can think of the sequence $\{U_k\}$ as the flow inter-arrival times, and of W_k as rate of the kth flow. The process $W(t)$ is then simply the instantaneous rate of a source. In [Taqqu and Levy, 1986], the authors consider the case where the inter-renewal times are heavy-tailed and the rewards have finite variance, and tackle the characterization of the aggregate process from a large number of sources at large time scales. They find that the limit process depends on which of the number of sources or of the time scale tends first toward infinity: it is a Lévy motion if the time scale goes to infinity first, whereas it is a fractional Brownian motion in the opposite case. Several improvements of these results have been added since, mainly to include more general assumptions on the reward distribution, and various limiting processes between Lévy stable motion and fractional Brownian motion are found [Levy and Taqqu, 1987, Levy and Taqqu, 2000] (see also [Doukhan et al., 2003]) in different limiting regimes. However, these improvements have received low attention from the network community, and we prefer to detail the results for a special case of the renewal models which is one of the most widely used model in the network community: the ON/OFF model. This model has been proposed in [Willinger et al., 1995, Willinger et al., 1997] to explain the self-similarity observed in aggregate network traffic, and the proofs of the results presented here appeared in the companion paper [Taqqu et al., 1997b]. The ON/OFF model is a renewal model where the rewards are strictly alternating between 1 and 0. The reward 1 corresponds to an ON period where a source is emitting with a constant rate (set to the normalized value 1), and the reward 0 corresponds to an OFF period (idle time or think time) where the source is not emitting. The ON (respectively OFF) periods are heavy-tailed distributed with tail index α_{ON} (respectively α_{OFF}). Both distributions are assumed to have a finite mean (*i.e.*, $\alpha_{\text{ON}}, \alpha_{\text{OFF}} > 1$), and at least one of these distributions is assumed to have an

infinite variance (*i.e.*, $\alpha_{\mathrm{ON}} < 2$ or $\alpha_{\mathrm{OFF}} < 2$). We denote by $W^{(n)}(t)$ the instantaneous rate of the source n at time t ($W^{(n)}(t) = 0$ or 1), and define the cumulative input process from N aggregated sources:

$$W_N(Tt) = \int_0^{Tt} \sum_{n=1}^{N} W^{(n)}(u)\mathrm{d}u.$$

Then, the following limit theorem holds (where the limit can be in the finite dimension distribution sense or the weak convergence sense, see [Taqqu et al., 1997b]):

$$\lim_{T\to\infty} \lim_{N\to\infty} \frac{W_N(Tt) - \mathbb{E}W_N(Tt)}{T^H N^{1/2}} = \sigma B_H(t) \tag{3.2}$$

and

$$\lim_{N\to\infty} \lim_{T\to\infty} \frac{W_N(Tt) - \mathbb{E}W_N(Tt)}{T^{1/\alpha_{\mathrm{H}}} N^{1/\alpha_{\mathrm{H}}}} = c\Lambda_{\alpha_{\mathrm{H}}}(t), \tag{3.3}$$

where $\alpha_{\mathrm{H}} = \min(\alpha_{\mathrm{ON}}, \alpha_{\mathrm{OFF}}) < 2$, H is as in relation (3.1), B_H is a fractional Brownian motion of Hurst parameter H, $\Lambda_{\alpha_{\mathrm{H}}}$ is a Lévy stable motion and σ and c are constants. If both the ON- and OFF-periods distributions have finite variance, then the limiting process is an ordinary Brownian motion (thus not long-range dependent) independently of the limits' order. Finally, in [Taqqu et al., 1997b], the authors also show that the result of equation (3.3) is valid for any finite number of sources, including one (removing the left limit). The result of equation (3.2) has been the most used in the last decade for network-traffic modeling. It states that the aggregate cumulative input process, properly rescaled, behaves at large time scales like a fractional Brownian motion, whose Hurst parameter satisfies relation (3.1). This implies two characteristics: its marginal is gaussian, and it is self-similar. Equivalently, one can say that the aggregate traffic (the increment process of the aggregate cumulative traffic) behaves like a fractional Gaussian noise: it is gaussian and long-range dependent. The limits' order in equation (3.2) is quite natural in network-traffic analysis: the traffic of a large number of sources is first aggregated, and then we look at this aggregate traffic at large time scales. However, when considering the reverse limit order, a strongly different process is found: the Lévy motion of equation (3.3). It has stable marginals (*i.e.*, non-gaussian and heavy-tailed), and its increments are not correlated. This surprising contrast between the two processes obtained in two limiting regimes raises several questions: how to understand the two limiting regimes? What happens if both N and T tend toward infinity together? In real systems with a finite number of sources and time scale, when is the Gaussian approximation valid? To understand the limiting regimes, one can think of the two following situations (see [Sarvotham et al., 2005]). If a small number of sources emit with a high rate, the aggregate process will likely be constituted of bursts, not stretched out in time. Since emission by each source is short, the aggregate traffic will not be correlated, and since the bursts have high rates, large values of the aggregate traffic are likely to happen so that the marginal will be heavy-tailed. This situation correspond to the limiting regime of equation (3.3). The time limit is achieved first because of the high rate of the sources ("time goes fast"). In the opposite situation, think of a large number of sources with rather slow emission rates. Because of the large number of sources, the aggregate process will likely be Gaussian, and because the emission rate is slow, there will be long flows introducing correlations at large time scales. This situation corresponds to the limiting regime of equation (3.2). The question "what happens if both N and T tend

toward infinity together?" is answered in [Mikosch et al., 2002]. The authors consider the number of sources as a function of the time scale T: $N(T)$ which tends to infinity when T tends to infinity. They introduce a slow growth condition under which the limiting regime of equation (3.3) is achieved and a fast growth condition under which the limiting regime of equation (3.2) is achieved. This is further elaborated in [Gaigalas and Kaj, 2003] where the authors show that a new process between Lévy motion and fBm is obtained in the intermediate case between the slow growth and fast growth conditions. Finally, for practical applications in real systems where the number of sources and time scale are finite, we need to know which limit regime is valid. It is suggested in [Taqqu et al., 1997b] that for a particular number of sources and time scale, to decide which regime is valid, it is sufficient to look at the marginal distribution: if it is Gaussian, then the process is fBm, and if it is stable then it is a Lévy motion. Conditions underlying the validity of the Gaussian approximation are given in [Kilpi and Norros, 2002] and imply a sufficient aggregation, both horizontal and vertical.

Infinite source Poisson models. In the simplest version of the infinite source Poisson models, flows arrive at the link as a Poisson process of rate λ, and transmit data at a fixed rate of 1 during a heavy-tailed random time of tail index α_{ON}. It is also known as the M/G/∞ model originally considered in [Cox, 1984]. Amazingly, in this model, the exact same results as for the ON/OFF model (equations (3.2) and (3.3)) hold true (see [Doukhan et al., 2003]) when replacing the number of sources N by the flow-arrival rate λ, and setting $\alpha_{\mathrm{H}} = \alpha_{\mathrm{ON}}$. It is shown in [Mikosch et al., 2002] that the same growth conditions also distinguish the two limiting regimes. Before further commenting on the relationship between the ON/OFF and infinite source Poisson models, we mention some of the numerous variants of the infinite source Poisson model, which all roughly rely on the same mechanism of a Poissonian arrival of some heavy-tailed flows and mainly differ on the way data is transmitted within a flow (see also a survey of infinite Poisson models in [Guerin et al., 2003], and the references in [Mikosch et al., 2002]). In [Kurtz, 1996], the authors consider an infinite source Poisson model with the general form of the "workload function" inside a flow and establish (for the first time) the Gaussian limit result. A similar model is considered in [Maulik and Resnick, 2003b] (where the "workload function" is called "transmission schedule") and the other limiting result is shown (the Lévy motion). In [Barakat et al., 2002], the Poisson shot-noise model is developed. This model is basically very similar to the two models previously mentioned. The "workload function" or "transmission schedule" is now called "shot", but "shots" still arrive according to a Poisson process. An application of this model is proposed in [Nguyen et al., 2004] where the shot shape is representative of the AIMD mechanism with Poisson losses. Even though it does not reproduce long-range dependence, we also mention the model proposed in [Ben Azzouna et al., 2004], where the M/G/∞ setting with Weibullian service times is used to model backbone traffic. Since the Weibullian distribution has finite variance, no long-range dependence property is found. In [D'Auria and Resnick, 2006], the authors consider a model where the rate within a flow is constant, but the rate of a flow is a random variable. The flow sizes are drawn at random according to a heavy-tailed distribution of tail index α_{SI}; and the rates are drawn at random independently of the sizes, according to any finite mean distribution. Flow duration (flow size divided by flow rate) is then heavy-tailed with $\alpha_{\mathrm{ON}} = \alpha_{\mathrm{SI}}$ and they find a long-range correlation structure of Hurst parameter H as in relation (3.1) with

$\alpha_H = \alpha_{SI} = \alpha_{ON}$. The last model we mention in this infinite Poisson model category is the Cluster Point Processes (CPP) model proposed in [Hohn et al., 2003]. In this model, the discrete approach of point processes is used. Flows are "clusters" of points and the flow arrival time is the time of the first packet of the cluster. The flow-arrival process is Poisson. The number of points in the cluster (the flow size) is heavy-tailed of tail index α_{SI}, and points (packets) within a cluster (flows) follow a renewal process with some inter-renewal distribution determining the mean flow rate. Based on results on point processes (see [Cox and Isham, 1980,Daley and Vere-Jones, 2003]), the authors calculate the aggregate traffic's spectrum and again find long-range dependence of Hurst parameter H as in relation (3.1) where $\alpha_H = \alpha_{SI}$. As in the model of [D'Auria and Resnick, 2006], the flow duration is this model is heavy-tailed of tail index $\alpha_{ON} = \alpha_{SI}$. As already observed in [Crovella and Bestavros, 1997], this might not hold true in real Internet traffic. For example, slow-start effects might imply higher rates for larger flows. In this case, flow rate is correlated to flow size and the tail indices of the flow-duration and the flow-size distributions are different. In [Hohn et al., 2003], the authors suggest as future work to introduce multi-class CPP, for example with a class for small flows (mice) having small rates, and another class for large flows (elephants) having high rates. Such a work, however, has not been done and the following question remains open:

- What happens if flow-duration and flow-size distributions have different tail indices? Which of these tail indices, if any, governs the long-range dependence of the aggregate traffic?

Relationships and differences between the ON/OFF model and the infinite source Poisson model. We have seen the ON/OFF and infinite source Poisson models yield very similar results, as far as the large-scale correlation structure is concerned (we limit this discussion to the Gaussian limiting regime of equation (3.2)). In both models, flow duration is heavy-tailed of index α_{ON}. However, the ON/OFF model is the only one capable of including some OFF times, and thus capable of reproducing long-range dependence that they can generate via their heavy-tailed distribution of index α_{OFF}. Actually, the basic difference between the ON/OFF model and the infinite source Poisson model lies in the flow-arrival process. While it is Poisson in the infinite source Poisson model, it is more complex in the ON/OFF case. Indeed, the flow inter-arrival time is the sum of the ON and OFF periods. Its distribution is then heavy-tailed with tail index $\min(\alpha_{ON}, \alpha_{OFF})$. The flow-arrival process then enters the framework of FRPs, and we have seen that the superimposition of a large number of such processes have almost exponential inter-arrival times, but it is not a renewal process. Results regarding the FRP show that the count process associated to the flow arrival is long-range dependent with Hurst parameter as in relation 3.1, where $\alpha_H = \min(\alpha_{ON}, \alpha_{OFF})$. In the case of the ON/OFF model, there can be two origins for the long-range dependence of the aggregate traffic: the correlation structure of the flow-arrival process, and the heavy-tailed ON periods. The results of the ON/OFF model show that whichever yields the larger correlations (*i.e.*, the larger Hurst exponent) dominates. If $\alpha_{ON} < \alpha_{OFF}$, both origins yield the same Hurst parameter for the aggregate traffic, governed by $\alpha_H = \alpha_{ON}$. If $\alpha_{OFF} < \alpha_{ON}$, the correlation structure of the flow-arrival process prevails at imposing long-range dependence with a Hurst parameter governed by $\alpha_H = \alpha_{OFF}$ In this last case, the heavy-tailedness of the ON times would yield a smaller Hurst parameter, and this potential origin then has no effect (taking exponentially

distributed ON-times would not change the Hurst parameter of the aggregate traffic). In the infinite source Poisson model, the flow-arrival process is uncorrelated, so that the effect of the heavy-tailed ON periods always dominates. Going back to papers [Hohn et al., 2002a, Hohn et al., 2003], where the authors, using semi-experiments, state that the long-range dependence of the flow-arrival process does not impact the long-range dependence of the aggregate traffic, we see that this conclusion was trace-driven, and not fully general: it is likely that in the trace studied in these papers, the long-range dependence of the flow-arrival process was not stronger than the one induced by the heavy-tailedness of the ON periods. Let us finally mention that the infinite source Poisson model can be seen as the limit of the ON/OFF model in some sense that we explain now. Consider the case where instead of applying the rescaling of equation (3.2), when adding sources (*i.e.*, when increasing N), we increase the mean OFF period μ_{OFF} so that the load remains constant. In this case, two consecutive flows of the same source are more and more separated in time when N increases. In the limit where N goes to infinity, two consecutive flows are separated by an infinite time, so that each source can be considered to emit only one flow. Since the sources are independent, the flow-arrival process is then Poisson and we recover the infinite source Poisson model (*c.f.* Proposition 2.1.3).

Other models. To conclude this section on models able to reproduce long-range dependence, we mention a few other models that are somewhat less related to our work. Several authors have proposed Markovian models able to reproduce pseudo self-similarity (*i.e.*, self-similarity in a finite scale range) [Robert and Le Boudec, 1997, Misra et al., 1998, Horváth and Telek, 2002, Khayari et al., 2004, Gong et al., 2005]. The advantage of these models is that they permit the use of well-known queueing-theory results in the Markovian framework and they can be of great interest in some practical applications. However, they hide the fundamental origin of asymptotic self-similarity: heavy-tails and more particularly infinite variance distributions (Markovian models are intimately linked to exponential distributions). We mention the existence of models based on dynamical systems [Erramilli et al., 2002]. These models use chaotic maps to reproduce a heavy-tailed ON time. Finally, even if it is not related to the rest, we mention the model proposed in [Rodolakis and Jacquet, 2005], where the authors show that long-lived TCP flows with random RTT drawn according to a heavy-tailed distribution generate long-range-dependent aggregate traffic. However, such a heavy-tailed distribution of the RTT has not been observed in the Internet.

3.1.2 Small scales

While there has been a large amount of work done to characterize large-scale properties of aggregate network traffic (see previous section), much less research work has been dedicated to the study of the small-scale properties of the traffic, which we briefly expose now. We first present two papers exploring the log diagram's behavior for a fixed order moment ($q = 2$) and focusing on the effect of pacing. We articulate the rest of the literature on the question: is network traffic multifractal at small scales? We first present papers giving a positive answer and the models proposed to explain this answer, and then we turn to papers giving a negative answer. We finally expose two mathematical models suggesting that TCP mechanisms can generate multifractality. An exhaustive survey of multifractal properties observed in network traffic can be found in [Veitch et al., 2005].

Pacing and burstiness at small scales

In [Jiang and Dovrolis, 2004b, Jiang and Dovrolis, 2005], the authors explore the log diagram of network traffic at small time scales, namely sub-RTT scales. They fix the order moment ($q = 2$) and thus do not tackle the question of the potential multifractality of the traffic. They focus on the effect of pacing and find that, as compared to TCP shaped traffic (where bursts are sent at the beginning of the RTT), paced traffic have much smaller *burstiness* at small scales. They characterize the burstiness by the absolute position of the log-diagram at a particular scale below RTT. Since they look at second-order statistics, their definition of burstiness is related to the variance of the process. They find that pacing has no effect on burstiness beyond the RTT scale. These studies clearly show that the small time scales behavior of the traffic mainly depends on the way packets are emitted within a flow (see also [Barakat et al., 2002]), and not on the flow-arrival structure, or the flow-size distribution as it was the case for large-scale properties.

Is aggregate network traffic multifractal at small scales?

The case for. The first evidence for multifractality at small scales in network traffic is reported is [Riedi and Lévy Véhel, 1997]. The authors study traffic captured at the gateway of LANs at Berkeley and CNET labs, and show that both the incoming traffic (WAN to LAN) and the outgoing traffic (LAN to WAN) are consistent with the multifractal property, even though their multifractal spectrum can be very different. Several processes are considered and yield the same conclusion: the "bytes per packet" process, the "time per packet" process and most importantly maybe the classical "bytes per time" process (*i.e.*, the count process of the aggregate traffic). Later, [Feldmann et al., 1998b] analyze WAN traffic from various locations between 1990 and 1997. The authors find that multifractality is not present in early traces, whereas this characteristic appears more pronounced in their last traces. They analyze the aggregate traffic averaged in time windows of size $\Delta = 10$ ms and find evidence for multifractality between 10 ms and roughly 500 ms. The same authors analyze in [Feldmann et al., 1998a, Feldmann et al., 1999a] (see also [Feldmann et al., 1999b]) WAN traffic from an ISP and from a corporate environment, acquired in 1997. In these traces, they find good agreement with a multifractal behavior on the aggregate traffic between 1 ms and 100 ms. To reproduce this small-scale multifractal behavior, several cascade models have been proposed [Riedi et al., 1999, Feldmann et al., 1998a, Feldmann et al., 1999b] (multiplicative cascades were at that time the only type of model known to generate controlled multifractal behavior). However, no direct relation between these models and meaningful network characteristics are given. In [Riedi et al., 1999], it is alluded that multiplicative effects might come from non-linear effect introduced by queueing behaviors. In [Feldmann et al., 1998a, Feldmann et al., 1999b], the authors suggest that the cascading behavior should be sought in the packet-transmission characteristics within a flow, but no more explanation is given. This is fully coherent with the theoretical work of [Maulik and Resnick, 2003b, Maulik and Resnick, 2003a] where the authors show that in an infinite source Poisson model, under fairly general conditions, the multifractal spectrum of the aggregate traffic is the same as the multifractal spectrum of the small-scale transmission process within one flow. Notwithstanding these meaningful contributions, they leave open the question of the fundamental origin of the multifractal property observed. A completely different interpretation of the multifractal property is

proposed in [Sarvotham et al., 2001, Sarvotham et al., 2005]. The authors analyze traffic traces acquired at University of Auckland, and propose to decompose the traffic in two components: the *alpha traffic*, constituted of a small number of sources emitting with a high rate; and the *beta traffic*, constituted of the rest (*i.e.*, a large number of sources emitting with moderate rates). As mentioned earlier, the *alpha traffic* is likely similar to the limit regime of equation (3.3) (Lévy process) whereas the *beta traffic* is likely similar to the limit regime of equation (3.2) (fBm). The authors of [Sarvotham et al., 2001, Sarvotham et al., 2005] then attribute the multifractal property of the total traffic to the *alpha traffic* part, since the Lévy process is known for its multifractal behavior [Jaffard, 1999]. However, this interpretation might again be trace-driven. Indeed, University of Auckland is in New Zealand, and traffic captured here is likely to have a bimodal distribution of the RTTs (this is actually shown in [Wang et al., 2002]): small RTTs for national communications and large RTTs for international communications. Small RTT connections thus constitute the alpha traffic (they achieve high rates), whereas large RTT connections constitute the beta traffic. Such a sharp distinction between alpha and beta traffic might not be possible in traffic from a less isolated country.

The case against. In parallel to this quest for the origin of the multifractal property, several articles even question its manifestation in real Internet traces. In [Taqqu et al., 1997a], the authors study the same traffic trace as in [Leland et al., 1993], which originally led to the discovery of self-similarity. They find a monofractal behavior for time scales larger that 100 ms, and they conclude that no multifractal model is necessary to describe the traffic. Based on Internet backbone traces, [Zhang et al., 2003, Ribeiro et al., 2005] also conclude that the traffic is consistent with monofractal models. The time scales studied in these papers is 1–100 ms, whereas RTTs experienced by the different connections are in the range 10 ms–1 s. To further precise the monofractal behavior observed, the authors analyze the small-scale regularity exponent and find that "high capacity" flows are able to create regularity exponents larger than 0.5 (*i.e.*, correlations at small scales), whereas traces without this type of flows have a regularity exponent of 0.5. The same conclusion is found in [Jiang and Dovrolis, 2004a]. Finally, in [Veitch et al., 2005], the authors review evidence of multifractality proposed in the past, and show that statistically poor estimation procedures have led to the wrong conclusion that the traffic is multifractal. Instead, they posit the use of the CPP model proposed in [Hohn et al., 2003], that predicts a monofractal behavior at small scales.

Two models of TCP generating multifractality. Outside of this debate about the potential presence of the multifractal property in real Internet traces, two articles propose TCP models that are consistent with usual multifractal properties [Barral and Lévy Véhel, 2004, Baccelli and Hong, 2002]. Differently to prior models mentioned above, these two models are based on a large number of long-lived TCP connections (rather than on aggregate-level ON/OFF structure). In [Barral and Lévy Véhel, 2004], the authors obtain a theoretical formula of the multifractal spectrum of the aggregate traffic with a fluid model of each TCP source. However, for tractability purposes, they assume random exponential losses and they have to consider an infinite number of sources, which introduce variability unrelated to TCP control mechanisms. In [Baccelli and Hong, 2002], the authors find multifractality in the aggregate traffic arising from a fluid AIMD model of many TCP

connections with congestion losses. However, their results are numerical and do not give analytical expressions relating the multifractal spectrum of the aggregate traffic to TCP mechanisms in various experimental conditions.

The debate that we briefly related above shows that the statistical properties of aggregate network traffic at small scales remain unclear: some authors observed multifractality at small scales in the aggregate traffic and attribute this property to the source traffic characteristics within a TCP flow. Some others question the evidence for multifractality. Some others finally propose models showing that the aggregation of a large number of TCP sources might lead to multifractality in the aggregate traffic. The question of the potential relation between TCP mechanisms and small scale multifractality of the aggregate traffic still remains unsatisfactorily answered. It seems however intimately related to the statistical properties of the TCP traffic at packet level that we will review in Section 3.3, but we first turn to the literature discussing potential impact of the statistical properties mentioned above (scaling laws and heavy-tails) on QoS.

3.2 Analysis of the Quality of Service

Impact of the traffic characteristics on QoS (in performance terms).

While many parameters can impact QoS (protocols, network topology, infrastructure, specific QoS mechanisms, etc.), we focus here only on the impact of the statistical properties mentioned above: scaling laws and heavy tails. We restrict ourselves to the single-buffer problem, and also do not relate considerable amount of work on scheduling policies. We divide the papers in two categories, based on the approach they use: open-loop analysis or closed-loop analysis.

3.2.1 Open-loop analysis (UDP): the impact of scaling laws

Since the discovery of long-range dependence and suspected multifractality in aggregate network traffic, a large amount of work has appeared on the evaluation of the potential consequences of these properties on QoS. The first category of papers that we present now use an open-loop method: they consider a long-range dependent (or multifractal) process feeding a buffer (most frequently of infinite size), and do not account for feedback. In this sense, they are more representative of the UDP case, even if some of these papers use real traces including TCP traffic. Consequently, these approaches consider only buffer QoS metrics, and they mainly concentrate on the delay or buffer occupancy.

Many of the results presented here can be found in the survey paper [Erramilli et al., 2002] and in the book [Park and Willinger, 2000]. As it was the case for statistical characterization of the traffic, much more research work have focused on the effect of long-range dependence on queueing behavior and fewer papers have considered the effect of small-scale scaling behavior and multifractality.

The impact of long-range dependence

Evaluation of the mean delay for infinite-size buffers. As a first QoS metric, a few papers have focused on the mean delay. In [Erramilli et al., 1996], the authors use

a real Ethernet trace exhibiting long-range dependence. To compare with the case where traffic is not long-range dependent, they use a *shuffled* version of the trace. These traces are used as the input of an infinite-size buffer, whose capacity can vary to achieve various utilization (load). They focus on the curve of the mean delay versus load, and conclude that beyond some *knee*, the mean delay increases much faster with the load for long-range-dependent input traffic. However, an interesting remark concludes their paper: this effect might not be observed with finite-size buffers, because the range of relevant time scales would then be limited. The same result (the mean delay increases faster with the load for long-range-dependent input traffic) is obtain in [Fiorini et al., 1997] based on a queueing-theory approach (the authors use a superposition of exponential distributions to emulate a heavy-tailed distribution: the TPT distribution proposed in [Greiner et al., 1999]).

Overflow probability for infinite-size buffers. The QoS metric that has been studied the most over the decade is probably the *overflow probability*: the probability that the buffer content exceeds a certain threshold, generally called B (*i.e.*, roughly the buffer-content distribution). The problem of determining this overflow probability for infinite-size buffers fed by long-range-dependent traffic has indeed generated a very large amount of theoretical works, which basically differ in the form of the input process (fBm, ON/OFF process, M/G/∞ process), and in the limiting regime that they consider (as for the models presented in Section 3.1.1, the limits order plays an important role, as well as the rescaling procedure). Most of them consider asymptotically large values of the threshold B (*i.e.*, the tail of the buffer-content distribution). In the seminal paper [Norros, 1994], the authors consider an infinite-size buffer fed by a fBm of Hurst parameter H. They show that the buffer-content marginal distribution $Q(B)$ is asymptotically of Weibull type:

$$\log Q(B) \underset{B\to\infty}{\sim} KB^{2(1-H)}, \tag{3.4}$$

where K is a slowly varying pre-factor. This Weibullian decrease implies a higher overflow probability than the case of a Markovian input process, where an exponential decrease is observed. It is likely to have stimulated the research activity on long-range dependence in network traffic, because it clearly predicts that long-range dependence can be a source of QoS degradation, at least as far as fBm is a good model for the input traffic and the buffer is large enough for the result to hold. Note that if $H = 1/2$, we retrieve the exponential decrease because the input process is no longer long-range dependent. Refinements of equation (3.4) are proposed in [Massoulié and Simonian, 1999, Narayan, 1998], where explicit formulae are given for pre-factor K. Following up this major contribution, many authors considered the case of an infinite buffer fed by an ON/OFF process. They consider N ON/OFF sources, with mean OFF time μ_{OFF}, feeding a buffer of capacity C. They are interested in the limit when N tends to infinity. To keep the load constant, some parameters have to be rescaled when N increases (note that all the studies consider a load less than 1 for stability). In [Simonian and Veitch, 1997], the authors rescale the capacity only: $C = Nc$, where c is a constant. They find a Weibullian decrease of the buffer-content distribution (as in [Norros, 1994]) if the ON and/or OFF period is heavy-tailed. In [Brichet et al., 1996], the capacity is kept constant, but μ_{OFF} is rescaled so as to keep the load constant (recall that this scaling corresponds to the limit to the M/G/∞ model). The decrease of the buffer-content distribution in this case is heavy-tailed if the ON periods are heavy-tailed (see also [Jelenković and Lazar, 1999] where similar results are found and

extended). The OFF-period distribution does not matter in this case since μ_{OFF} goes to infinity. A third rescaling scheme is used in [Mandjes and Borst, 2000, Mandjes and Boots, 2001], where both the capacity and the threshold are rescaled: $C = Nc, B = Nb$, where b and c are constants. For large b values, the authors find that, independently of the OFF-times distribution, the buffer-content distribution decay is heavy-tailed if the ON-time distribution is heavy-tailed whereas it is exponential if the ON-time distribution is not heavy-tailed. An interesting advantage of this third rescaling scheme is that it allows the authors to derive results in the case of small values of b (*i.e.*, small overflow threshold). They find that in this case, the overflow probability depends on the ON and OFF distribution only through their mean [Mandjes and Kim, 2001, Mandjes and Boots, 2001]. This insensitivity result confirms previous findings of [Ryu and Elwalid, 1996, Neidhardt and Wang, 1998]: for small thresholds, long-term correlations do not impact the overflow probability, even if the buffer has an infinite size. These findings were based on the approximation of the critical time scale: the most likely amount of time it takes for the queue to fill up beyond the given threshold. While the critical time scale approximation is a powerful theoretical tool, it is impractical because its computation requires statistics (marginal distribution) of the traffic at *all* time scales. Computing it is then easy for parsimonious models like fBm, but becomes much more difficult for more general types of traffic or if one want to compute it directly from measurement. To overcome this drawbacks, some further approximations are proposed in [Ribeiro et al., 2006b].

Finite-size buffers. As remarked in [Heyman and Lakshman, 1996], it is not equivalent to study an infinite-size buffer, and look at the probability that the buffer content exceed a threshold B; or to study a finite buffer of size B and look at the drop probability. Indeed, dropped packets reduce the number of packets that get into the buffer and thus shorten the busy periods. The overflow probability in the infinity buffer is then an upper bound of the loss probability in the finite buffer. However, very few papers have considered the case of a finite-size buffer. In [Heyman and Lakshman, 1996], the authors propose a discrete model based on *cells* for variable bitrate (VBR) traffic. They use the critical time-scale approximation to show again that long-term correlations do not impact the loss probability. They conclude that Markov models are sufficient to study the loss probability for small buffers. In [Grossglauser and Bolot, 1999], the authors introduce a non-tractable model and use numerical simulations to come to the same conclusion. Finally, theoretical large deviation results for the case of finite-size buffers are obtained in [Jelenković and Momčilović, 2003], but their validity holds for very large buffers only.

The gain of multiplexing and the importance of marginal distributions. Although it is not related to the question of the impact of long-range dependence on QoS, we mention here two important remarks recurrently appearing in the papers mentioned above. Firstly, and independently of the presence of long-range dependence, there is a gain of multiplexing. Indeed, for the same mean aggregate traffic, the more sources there are, the smaller the variance (at each scale). This effect can be seen for example in a reduction of the pre-factor of equation (3.4), or equivalently, on the absolute position of the log diagram (lower for a smaller variance, independently of its slope). This effect is mentioned for example in [Grossglauser and Bolot, 1999, Heyman and Lakshman, 1996, Erramilli et al., 2002], but many other papers make the same remark. The second remark is related to

the first one: marginal distributions are very important for the evaluation of QoS. This is very clear when considering the critical time-scale approximation, which depends on the marginal distributions of the input traffic at each time scale, and in [Ribeiro et al., 2006b], where the approximations of the critical time scale proposed depend on the distributions at a subset of time scales. This second remark is clearly related to the first one in the case of Gaussian distributions, where the variance is one of the two parameters of the distribution. However, the authors of [Grossglauser and Bolot, 1999] stress that the mean of the distribution is also an extremely important parameter, that should be properly taken into account.

The impact of small-scale scaling behavior

We now turn to the few papers discussing the impact of small-scale statistical properties of the aggregate traffic on QoS. The same trace-driven method as in [Erramilli et al., 1996] is used in [Erramilli et al., 2000], where the authors extend the study to the effect of small-scale multifractality (to suppress multifractality from the real trace, the authors uniformly distribute packets in a small time windows of 512 ms). They find that, while long-range dependence has a strong impact at high loads, like in the previous paper [Erramilli et al., 1996], for small loads it is the multifractal property that is likely to degrade the performance (mean and maximal delay here). Finally, in the previously mentioned papers [Jiang and Dovrolis, 2004b, Jiang and Dovrolis, 2005], the authors also study the impact of small-scale properties on the delay, focusing on the effect of pacing. They consider the log-diagram for the second-order moment only (*i.e.*, the variance) and are not interested in potential multifractality issues. They show that pacing reduces the variance of the aggregate traffic at sub-RTT scales. Using this aggregate traffic as the input of an infinite-size buffer of variable capacity, they find that paced traffic generates smaller mean delay for low utilization (below 70%) only. They suggest that for higher utilization, long-range dependence effects would predominate. These results are coherent with the results presented above, in particular those for small-threshold overflow probability and finite buffers: for small loads, the buffer content is *globally small*, so that the critical time scale is small and the long-term correlation has no impact, whereas higher loads naturally increase the buffer utilization and the critical time scales so that long-range dependence becomes an important factor.

3.2.2 Closed-loop analysis (TCP): the impact of heavy-tailed distributions

The papers presented in the previous section use an open-loop method to evaluate buffer QoS metrics, such as delay. While considerable insights are gained from these studies, they are more suitable to model UDP traffic, and the fact that they do not consider feedback make their direct use difficult for TCP traffic [Arvidsson and Karlsson, 1999, Joo et al., 2001]. For TCP traffic, the approach is fundamentally different because the QoS observed at the buffer (loss) directly impacts the emission process by the sources and then the aggregate traffic. It is no longer possible to assume *a priori* specific properties of the aggregate traffic and study separately the impact of these properties on QoS, as it was the case for UDP traffic. The question of interest is then to characterize QoS given the flow-level properties of the sources (*i.e.*, flow-size distribution and OFF-time distribution). In the case of TCP, the QoS metrics of interest are at the buffer level (delay, loss, throughput) but also at the receiver (completion time). While many parameters, in particular the protocol,

can impact QoS [Dukkipati and McKeown, 2006], we focus our presentation on the impact of a long-tailed flow-size distribution on QoS.

An empirical study of the impact of α on QoS. A first and seminal empirical study of the effect of heavy-tailed flow-size distribution on QoS appeared in [Park et al., 1997]. The authors use ns simulations with 32 ON/OFF sources where the flow-size distribution is a pure Pareto distribution of fixed mean and of varying tail index α; and the OFF-times distribution is exponential. They find that a small value of α systematically degrades the QoS metrics: loss, mean delay and throughput, but that this effect is smaller for very small values of the buffer size (3 packets). These results however might be affected by specific deterministic effects of the simulator at small time scales, especially for small buffer sizes. Moreover, the pure Pareto distribution only has two parameters: α and the lower cutoff k (fixing the mean for a specified value of α). Thus, varying the α parameter for a fixed mean not only changes the tail of the distribution but also its body, and does not permit to assess the isolated effect of the tail of the distribution on QoS parameters.

Insensitivity results. Apart from the paper mentioned above, the largest part of the literature on the impact of heavy-tailed distributions on QoS has concluded to *insensitivity*, *i.e.*, there is no impact of heavy-tailed distributions for TCP traffic [Arvidsson and Karlsson, 1999, Heyman et al., 1997, Massoulié and Roberts, 2000, Ben Fredj et al., 2001, Roberts, 2004]. Basically, the argument supporting this result is that QoS metrics depend only on the distribution of active flows at an instant, which depends only on the mean of the ON-times distribution as long as the flow-arrival process is Poisson. More precisely, the authors of [Ben Fredj et al., 2001, Roberts, 2004] show that if the flow arrival is Poisson, then the first-order QoS statistics depend only on the mean of the flow-size distribution. The QoS metrics considered are the mean throughput and the completion time (which is shown to be proportional to flow size). Without questioning the Poisson flow-arrival assumption (which is certainly justified at small time scales for a large number of sources, but see also [Schroeder et al., 2006, Prasad and Dovrolis, 2008]), their results also use another simplification: equal bandwidth sharing. They assume that bandwidth is equally shared between the active connections and the authors of [Ben Fredj et al., 2001] note that their result might not hold if "the rate limit is generally different for every flow and can vary during the transfer". These results might then fall down if transient behaviors (like for example slow start) predominate, or if different TCP variants coexist.

Fixed-point approaches. Due to the closed-loop characteristic introduced by TCP's feedback mechanisms, fixed point approaches have naturally been proposed. The first one appeared in [Casetti and Meo, 2000]. The authors model separately the sources' behavior and the network. The behavior of a source is modeled by a complex Markov chain with two "big states", ON and OFF. Within the big state ON, they introduce many states representing the evolution of the source's congestion window. The parameter of this Markov chain is the loss probability. To model the network, they use classical queueing models, whose output is the loss probability. They use a fixed-point approach to determine the loss probability in the *operating point*. Extensions of this method to include specific mechanisms of TCP Vegas are considered in [Wierman et al., 2003b, Wierman et al., 2003a]. However, this fundamentally Markovian approach is limited to exponential ON and OFF

distributions. To overcome this limitation, a similar approach is used in [Garetto and Towsley, 2003, Garetto and Towsley, 2008] where the authors use a superposition of exponential distributions to emulate a heavy-tailed distribution, but no evaluation of the impact of the tail parameter is proposed. Moreover, these models based on Markov chains with many states to account for the numerous specificities of TCP have a lot of parameters (especially if several such chains have to be superposed to account for a heavy-tailed distribution). They allow for numerical results only, and extracting qualitative behavior is difficult. Their large complexity also reduces their trustability from the viewpoint of the practically-interested community, which often prefers experimental evaluations.

In the case of the UDP protocol, most of the work on the impact of scaling laws/heavy-tails have been done in the framework of infinite queues and give asymptotical results either for very small or very large buffers. On the other side, in the case of the TCP protocol, proposed models have such a complexity that it is difficult to extract qualitative tendencies. Finally, in both cases, the following question have received partial answers only, which need comprehensive empirical studies to be completed:

- In realistic conditions implying finite buffer sizes, what is the impact of heavy-tailed flow-size distributions on QoS (in performance terms)? More precisely, on which QoS metric?

3.3 Source-traffic characteristics at packet level

TCP traffic at packet scale and TCP throughput predictability.

In Section 3.1.2, we have mentioned the literature on the statistical properties of the aggregate traffic at small time scales, and have concluded that the potential relationship between TCP mechanisms and aggregate traffic multifractality remains unclear. We have seen however, that many authors invoke characteristics of the source emission process within a flow to explain the presence of multifractal properties in the aggregate traffic, so that we now review the literature related to this question. With the UDP protocol, the question of source traffic characterization has limited interest because packets are regularly emitted within a flow. In TCP, packets are emitted by bursts at the beginning of each RTT. The number of packets in a burst is called the congestion window, directly related to the QoS parameter *throughput*. A large amount of work have discussed the evolution of this congestion window under various external conditions. Most of these works have been done in the framework of *long-lived TCP connections*, which does not include the flow-level structure (ON/OFF) of a source. This framework represents well the packet-emission process within a long flow. Surveys of the results briefly presented here can be found in [Barakat, 2001, Erramilli et al., 2002, Budhiraja et al., 2004]. As we shall see, most of these results deal with the first-order statistics (mean) of the congestion window and do not discuss potential multifractality. For this reason, our presentation is quite short.

Recall that we concentrate on the single bottleneck buffer, and we do not develop the large amount of work that appeared on the case of TCP connections going through a network of queues, and the stability problems that this generates [Kelly et al., 1998, Bonald, 1998, Gibbens et al., 2000, Kelly, 1999, Roughan et al., 2001, Baccelli and Bonald, 1999, Baccelli and Hong, 2000, Baccelli and Hong, 2005].

One singleTCP connection

We first present the works considering only one TCP connection. Under various loss-process assumptions, with various methods (Markov, fluid model), and with various level of complexity (e.g., introducing or not TCP mechanisms like timeout and slow start) all of them are basically trying to evaluate the mean throughput achieved by the TCP connection (*i.e.*, the mean congestion window).

A first series of papers assumes Bernoulli losses (*i.e.*, independent packet losses). This is the case of [Mathis et al., 1997, Lakshman and Madhow, 1997, Padhye et al., 1998] where the authors show, using a simple Markov chain to model the evolution of the congestion window, that the mean congestion window goes as the inverse square root of the loss probability p:

$$\mathbb{E}\{cwnd\} \underset{p\to 0}{\sim} \frac{K}{\sqrt{p}}, \tag{3.5}$$

where K is a constant. This is the well-known *square root formula*, and applies for small loss rates. In [Padhye et al., 1998, Padhye et al., 2000], the authors extend the model to account for the timeout mechanism, and obtain a more complex formula that applies over a wider range of loss rates, but nevertheless reduce to the formula (3.5) for very small loss rates. Many other extensions of this models have also been proposed in [Fortin-Parisi and Sericola, 2004] (to include the slow start mechanism), [Kaj and Olsén, 2001] (to include various mechanisms like maximal advertisement windows, fast retransmit and ack threshold) and [Sikdar et al., 2003, Kumar, 1998] (to include TCP variants other than RENO). They also assume Bernoulli losses and focus on the evaluation of the mean congestion window. Instead of the discrete Markov chain, several authors have used fluid models of TCP. This is the case for example of [Misra et al., 1999] where the authors assume that the loss process is Poisson and obtain the mean throughput by solving a "Poisson counter-driven stochastic differential equation" (see also [Budhiraja et al., 2004] and the reference therein).

Several propositions have appeared to handle more general loss processes than Bernoulli. In [Kamal, 2004], the authors use a Markov chain to model the evolution of the congestion window of one TCP connection, and a "discrete batch Markov arrival process" is used to model the rest of the traffic. In [Altman et al., 2000, Altman et al., 2005], the authors concentrate on the congestion window just before a loss and they recover the mean throughput via Palm calculus, using a stationary ergodic loss process. They find a more general formula than (3.5) for the mean throughput. This line of thought is extended in [Blanc et al., 2009b, Blanc et al., 2009a] to different TCP variants, but with a Bernoulli loss assumption.

We mention two other types of extensions of the fluid models for long-lived TCP sources. In [Carofiglio et al., 2007], the authors propose various extensions of fluid models to take into account small-scale variability. In [Baccelli and McDonald, 2006], the square root formulae is extended to ON/OFF sources with Bernoulli losses. A further extensionis proposed in [Baccelli et al., 2007] to include rate-dependent losses.

Multiple TCP connections

The papers presented above were considering a single TCP connection, and trying to evaluate the mean congestion window. A few papers have tackled the modeling of several competing TCP connections. In this case, two important new notions have to be considered: fairness and synchronization. Roughly, synchronization represents the tendency that

several sources experience a loss event at the same time, and fairness is the general term used to describe how fairly several sources share the available capacity of a link. In [Aithellal et al., 1997], the authors consider the case of two competing TCP connections and use a fluid model to evaluate their mean throughput. In [Hurley et al., 1999], multiple TCP connections receiving feedback depending on their sending rate are studied. Based on a fluid model and resolution of ordinary differential equations, first-order statistics of the fairness are obtained. In [Qiu et al., 2001], the impact of synchronization on the mean throughput of many TCP connections is experimentally studied. Finally, in [Baccelli and Hong, 2002], the authors use a fluid model of N TCP sources reproducing the AIMD behavior. Using products of matrices, they obtain the steady state distribution from which they deduce, in addition to the first-order statistics of the throughput of each connection, the autocorrelation function and the covariance matrix. They also derive results on fairness showing that instantaneous throughputs of each connection can exhibit a large dispersion (with some heavy-tailed dispersion index). As a possible mean to describe the variability of TCP traffic, and characterize it beyond its mean, we mentioned at the end of Section 3.1, the authors of [Baccelli and Hong, 2002] also show that aggregated traffic from a large number of long-lived TCP sources following their fluid AIMD model exhibits small-scale multifractality. However, this result is numerically supported only. Moreover, it does not link the multifractal property of the aggregate traffic to the TCP mechanisms governing the behavior of each sources. This is also the case of [Barral and Lévy Véhel, 2004] mentioned earlier (Section 3.1.2).

None of the papers mentioned above have given conclusive answers regarding the potential multifractal characteristic of TCP traffic, that could explain the multifractality of aggregate traffic. The models proposed are often restricted to prediction of first order statistics of TCP throughput (the mean), and the following question of primary importance remains open:

- Do TCP mechanisms generate multifractal scaling laws in the packet-level source traffic? What is their relation with TCP throughput prediction (beyond the mean)?

Conclusion

We summarize here the open questions that we have pointed out in this chapter, and on which we concentrate to propose answers in the rest of the manuscript. For the sake of completeness, we have included in this summary the questions related to the experimental platform and to the flow size distribution tail index estimation for which the state of the art will be reviewed in the dedicated chapters (Chapters 4 and 6).

Experimental platform reproducing realistic conditions.
⟶ **Chapter 4**

Aggregate network traffic characterization.

- To what extend is Taqqu's relation valid in real network conditions? What are the scales involved? Does the protocol play a role? In which situation (loss-free, lossy link, congestion)?
⟶ **Chapter 5, Sections 5.1 and 5.2**

- What happens if flow-duration and flow-size distributions have different tail indices? Which of these tail indices, if any, governs the long-range dependence of the aggregate traffic?
⟶ **Chapter 5, Section 5.3**

Source traffic characterization at flow scale.

- How to estimate the flow-size distribution's tail index from packet-sampled data?
⟶ **Chapter 6**

Impact of the traffic characteristics on QoS (in performance terms).

- In realistic conditions implying finite buffer sizes, what is the impact of heavy-tailed flow-size distributions on QoS (in performance terms)? More precisely, on which QoS metric?
⟶ **Chapter 7**

TCP traffic at packet scale and TCP throughput predictability.

- Do TCP mechanisms generate multifractal scaling laws in the packet-level source traffic? What is their relation with TCP throughput prediction (beyond the mean)?
⟶ **Chapter 8**

Chapter 4

Experimental and numerical aspects

The results of this chapter have been published in [4] [1] [10,9], and presented as a demonstration in [12].

Experimental platform reproducing realistic conditions.

Introduction

Chapter 1 introduced the necessity of an experimental approach of questions related to the analysis of network traffic and QoS, in particular due to the complexity introduced by protocol TCP. In this chapter, we present the metrology platform that we developed to perform large-scale experimental studies of network traffic and QoS in realistic conditions, including a large number of controlled sources, high-speed dedicated links and high-performance traffic-measurement tools.

We first discuss some existing solutions, for the experimental platform and for the measurement tools. Then we present our solution to perform controlled experiments in realistic conditions (Section 4.2). It is based on the large-scale fully controllable testbed *Grid5000* (Section 4.2.1). In Section 4.2.1, we present our experiments, and the tool we developed to automate them and handle the logs collection: *TranSim*. Section 4.2.2 is dedicated to the presentation of our solution for packet-level traffic capture at up to 10 Gbps: *MetroFlux*. While other solutions could have been possible (based on DAG cards for example), our tool is based on the FPGA device *GtrcNet*, which perform traffic capture and various other functions, such as sampling, traffic generation, latency emulation, etc.

The versatility of our capture tool *MetroFlux* allows us to also plug it on real links. This way, we obtain real traffic traces from such links that complement controlled experiments performed on Grid5000. Such traces are presented in Section 4.3.

While studies of network traffic involving the TCP protocol necessitate an experimental approach performed on a real network, UDP traffic is less complex, and interesting studies can be performed using simple numerical simulations to complement the experimental approach. To perform such simulations, we develop matlab routines that we present in Section 4.4.

We conclude and provide a summary of the controlled experiments, traces and simulations used in the rest of the manuscript, at the end of this chapter.

4.1 Existing solutions for empirical network-traffic analysis

This section presents some existing possibilities to empirically tackle questions related to network traffic and QoS analysis. This survey is non exhaustive, and we do not claim that there does not exist similar platforms that could have fulfilled our requirements under slight alteration.

4.1.1 Experimental platforms and simulators

Real Internet traces. To analyze statistical characteristics of network traffic, a first and natural method relies on the study of real Internet traces. To this end, many Internet traces, acquired at various points of the Internet backbone are publicly available, in particular from the National Laboratory for Applied Network Research project (NLANR). However, while the analysis of real Internet traces is essential to characterize the main features and trend of recent traffic in realistic conditions, it has limitations. Indeed, Internet traffic parameters (*e.g.*, the load, the flow-size distribution) are not controlled. It is thus impossible to change these parameters and study their effect on the system. Moreover, Internet traffic presents cyclic trends related to the daily activity periods, that prevents the observation of long stationary traces.

Simulators. To overcome these limitations, simulators have been massively used in the research community. The most used network simulator in empirical analysis of network traffic and QoS is *ns* [22]. It is a discrete event simulator, providing substantial support for simulation of TCP, routing, and multicast protocols over wired and wireless (local and satellite) networks. However, simulators also have restrictions. 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. Simulators also have a determinism issue: they cannot take into account the small-scale randomness inherent to real systems.

Experimental platforms. Large-scale experimental facilities are alternatives that may overcome the limitations of both the Internet and simulators, as they permit controlling network parameters and traffic generation, including statistics and stationarity issues. A lot of initiatives appeared to propose such solutions, with different objectives (GENI, FIRE, etc.). We expose here only a few of them, among the most famous. Emulab [White et al., 2002] is a network experimental facility where network protocols and services 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, users specify an arbitrary network topology, having a controllable, predictable, and reproducible environment. They have full root access on PC nodes, and they can run the operating system of their choice. However, the core network's equipments and links are emulated. The PlanetLab testbed [Bavier et al., 2004] currently consists of about 1000 PCs on about 500 sites (mostly hosted by universities with about 2 PCs on each site) 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. However,

the non-isolated links used in PlanetLab do not permit an isolated study of network traffic generated in controlled situations. Moreover, it does not easily permit the inclusion of measurement tools at chosen points of the network. The OneLab project in Europe provides an open federated laboratory, built on PlanetLab Europe, which aims at supporting network research for the future Internet. A versatile monitoring software (COMO) was notably developed within this project (a new version is still under development).

In our experimental platform, we use the *Grid5000* testbed. As we shall see in Section 4.2.1, this experimental facility fulfill our requirements: a large number of fully controlled machines (around 5000 machines on 10 sites) and high-speed dedicated links (10 Gbps). Moreover, because we are members of the *Grid5000* project, slight alterations of the configuration were possible to adapt it to the specific requirements of our experiments. In particular, we could easily include traffic-measurement tools at key points of the network.

4.1.2 High-speed traffic capture

In addition to the capability of generating traffic from a large number of controlled sources in a realistic environment, a complete characterization of the traffic requires its capture at key points of the network.

A few studies have already focused on the feasibilty of network-traffic capture at speeds as high as 10 Gbps, using commodity hardware [Schneider et al., 2007,Anderson and Arlitt, 2006]. Most of these studies and of the few solutions available on the market to perform traffic capture (e.g., products from Endace, Solera Networks, or GigaStor) focus on full traffic capture, whereas our requirements are more focused on the extraction and capture of packet headers, which are sufficient for our needs. Besides, a lot of the solutions available today are designed and marketed as intrustion-detection or -prevention systems, and focus on real-time deep-packet inspection rather than traffic capture.

Moreover, our requirements also include the capability of capturing headers with no loss even in the worst case of very small packets arriving at the maximal load. In this worst case, the capture system has to cope with a data bandwidth which is almost as high as if it was storing the full packets instead of the headers, so that the technical problems are quite similar:

- The capture server has to handle a high number of packets per second, which will in turn trigger a high number of interrupts per second, if using the legacy network stack.
- The data path in the capture system has to be as short and direct as possible, and keep data copy to a minimum.
- The disk storage has to be both very large and very fast.

In [Schneider et al., 2007], the authors come to the conclusion that commodity hardware is unable to fulfill all these requirements and they propose to distribute the capture among several servers. DAG cards, from Endace, which are quite popular in the research community, address the first two problems by optimizing the data path in the host system and providing an appropriate API. The authors in [Anderson and Arlitt, 2006] solve the first two problems by using a state-of-the-art (at the time of writing) high-performance server, carefully tweaked for the task, and custom software optimizing the capture. They solve the storage problem by using a very powerful raid array. They also use a special dataseries file format for the storage.

Such solutions, based on DAG cards, for example, could have been acceptable to fulfill our traffic capture requirements. However, we chose to use another solution, based on FPGA device *GtrcNet*, that in addition offers various other features and large reconfigurability.

4.2 Controlled experiments

4.2.1 Grid5000: a large-scale testbed

Grid5000 overview

Grid5000 [15], is a 5000 CPUs nation-wide Grid infrastructure for research in Grid computing [Bolze et al., 2006], providing a scientific tool for computer scientists similar to the large-scale instruments used by physicists, astronomers, and biologists. It is a research tool featuring 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 are involved and 9 sites host at least one cluster, each made of about 500 cores. A dedicated private optical networking infrastructure, provided by RENATER, the French National Research and Education Network, interconnects the Grid5000 sites. Two international interconnections are also available: one at 10 Gbps interconnecting Grid5000 with DAS3 in the Netherlands, and one at 1 Gbps with Naregi in Japan. A new interconnection with a site in Porto Alegre (Brazil) is also available since July 2009. In the Grid5000 platform, the network backbone is composed of private 10 Gbps Ethernet links connected to a DWDM core with dedicated 10 Gbps lambdas (see Figure 4.1).

Grid5000 offers to every user full control of the requested experimental resources. It uses dedicated network links between sites, allows users to reserve the same set of resources across successive experiments, to run their experiments in dedicated nodes (obtained by reservation) and it allows users to install and to run their own experimental condition injectors and measurement software. Grid5000 exposes two tools to implement these features: OAR is 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 enables researchers to run successive experiments reproducing the exact experimental conditions several times, an almost impossible task with shared and uncontrolled networks. This also ensures large-duration observation windows under stationary conditions – something that cannot be achieved on the Internet. As a private testbed from which we are members, Grid5000 makes installing experimental hardware, such as the traffic capture instrument at representative traffic aggregation points, quite easy..

In terms of monitoring tools, users have access to the RENATER Monitoring tool, the Grid5000 infrastructure description, the Nagios tool, Ganglia and SFlow (only at lyon). The RENATER Monitoring tool provides a view of all connections on the Grid5000 backbone. With it, users can see the availability and the usage of the global links. The Grid5000

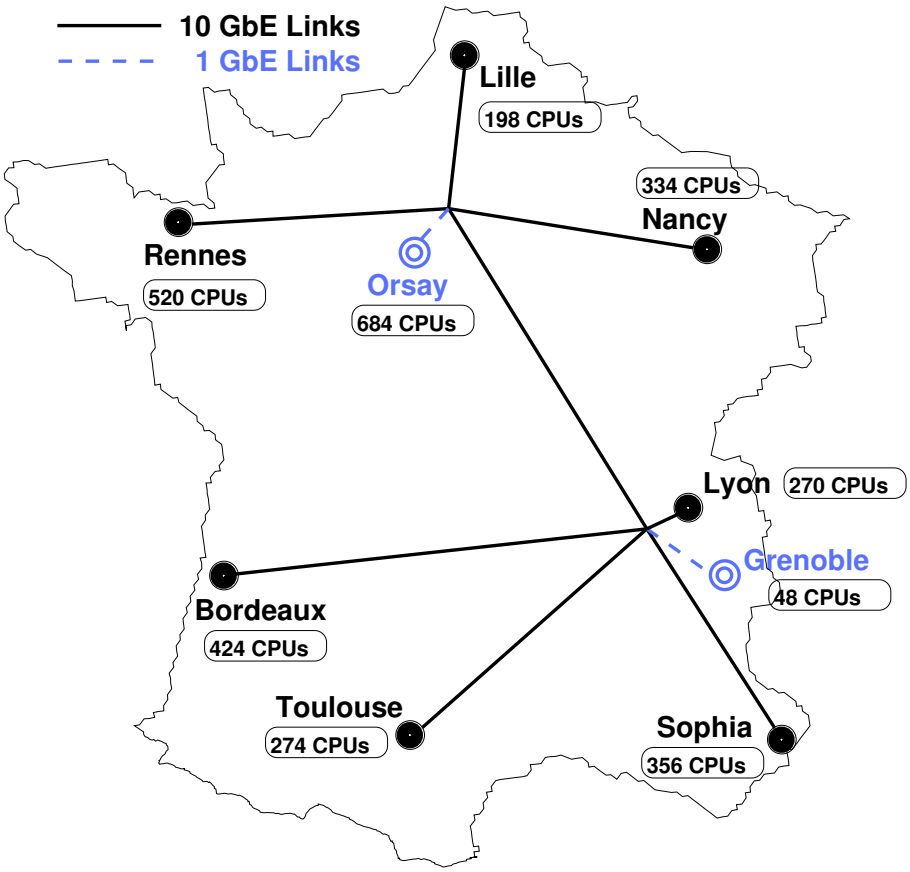

Figure 4.1: Grid5000 backbone

infrastructure's description gives all the information about the different Grid5000 platforms. It shows, graphically, on all sites their internal connections, that means, what kind of network hardware is used and what are the different links between them. The Nagios tool is used to warn the administrators about connectivity and services problems. Ganglia provides users with the rate usage per node and other information, not necessarily network specific, such as the available disk space, the cpu usage, and so on. The Sflow tool is an experimental monitoring tool which provides a sampling of the network traffic. The common NetFlow tool providing flow-level information is also available on several Grid5000 sites.

Description of our experiments

We now describe more precisely the experiments we performed, and the tool we developed to automate them: *TranSim*.

Topology To obtain the single bottleneck topology described in Chapter 1 (Figure 1.1), we always use the site of Lyon for the sources and one of the other sites of Grid5000 for the destinations. The bottleneck is always at the site of Lyon. In all our controlled experiments, we use a bottleneck of 1 Gbps. Prior to January 2008, the uplink of the Lyon site was at 1 Gbps, and the traffic was monitored at the principal router of Lyon, as shown in Figure 4.2. This topology is used in the experimental validation of Taqqu's relation, presented in Section 5.1. In this configuration, 140 machines were available at the Lyon site, but only 100 were actually used in our experiments. Each source has a maximal emission rate of 1 Gbps, but as we shall see, we can use different rate-limitation mechanisms to limit this source emission rate, in order, for example, to avoid congestion. The destination site used was Rennes, and the minimal RTT between Lyon and Rennes was then 12 ms.

After January 2008, the uplink of the Lyon site was upgraded to 10 Gbps, and the topology that we use in all our experiments (except those of 5.1) is shown on Figure 4.3. Up to 46 machines are monitored at a 1 Gbps switch in Lyon. One of these machines is also monitored before entering the switch, so that for QoS analysis, we can compare the input and output traffic of this source. Each of these 46 machines can have a different source behavior. For example, some of them can use the UDP protocol, and some other the TCP protocol with different TCP variants, etc. In addition, we can use the 94 remaining machines of the Lyon site to impose a cross traffic in order to create congestion at the 10 Gbps output of the Lyon site. These 94 machines will be used only to emulate a lossy link at the output of the 1 Gbps switch, in Section 5.2.1. The destination site in the experiments using the topology of Figure 4.3 is usually Rennes or Nancy, leading to a minimal RTT of 12 ms or 10 ms respectively. Note finally that the second topology of Figure 4.3 has the advantage that we can vary the switch's output buffer size, between 96 and 4096 packets (this would not have been possible with the first topology of Figure 4.2, because we cannot change the buffer size of the principal router of the Lyon site).

In all our experiments, we use "tap"s to duplicate the traffic to be captured. This way, we avoid potential delays introduced in the switch-mirroring operation.

Transfer generation Typically, in our experiments, a large number of sources emit "flows" following an ON/OFF scheme where ON and OFF periods are random variables.

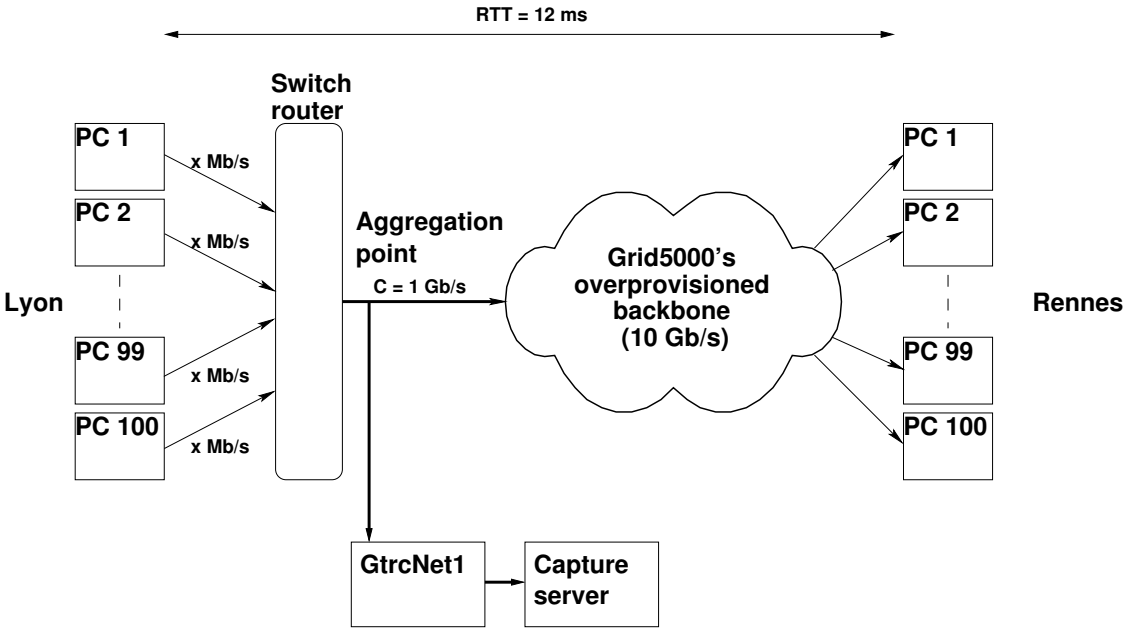

Figure 4.2: Topology1, used in the experiments before January 2008 (experiments of Section 5.1).

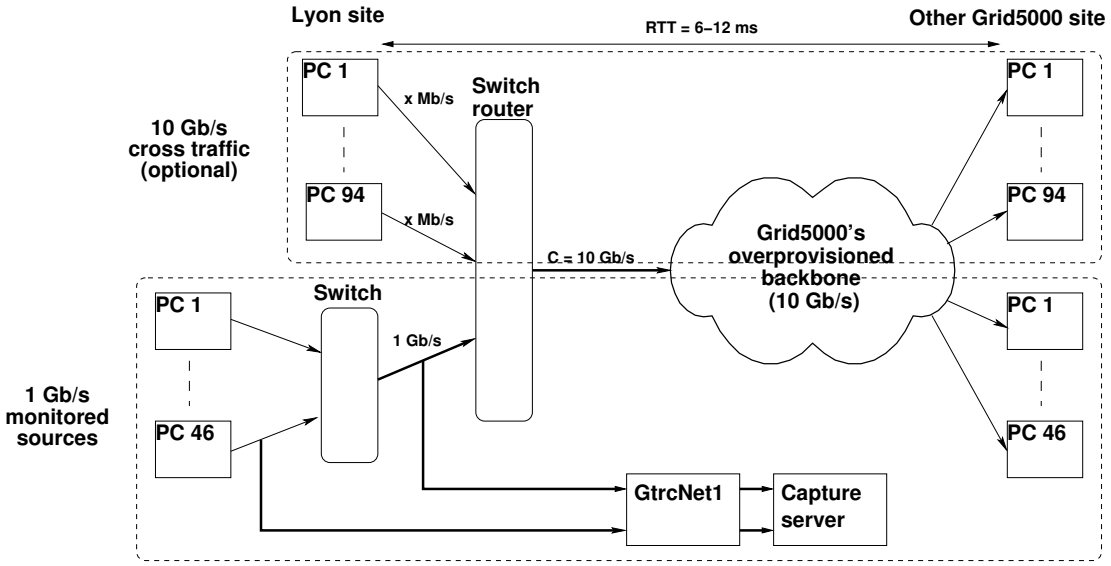

Figure 4.3: Topology2, used in the experiments after January 2008.

	description
Client nodes	Sun Fire V20z (bi-opteron)
Kernel	GNU/Linux 2.6.18.3
iperf version	2.0.2

Table 4.1: Characteristics of the sources (nodes of the Lyon site).

It is thus impossible to start these transfers manually, and so we developed tool *TranSim* to automate the transfers and log collection.

TranSim automates each step of an experiment:

- reservation of the required nodes;
- deployment of the environment on each node;
- launch the measurement tools (capture, web100);
- start the sequence of transfers (with *iperf*) and OFF times (with the *sleep* command) for each source;
- stop the experiment after a predefined time;
- stop the capture;
- log collection (web100, iperf, netstat).

TranSim also allows us to have several sources located on the same machine (or node), which are independent at flow scale. This way, if we use two independent sources on each node, we can have up to 92 independent monitored sources in the topology 2 of Figure 4.3. Two sources located on the same node have independent ON/OFF schemes, but share the same network interface.

TCP and UDP transfers are realized using *iperf* [18]: a traffic-generation tool that allows users to tune the different TCP and UDP parameters so as to evaluate their impact on network performance. Among the numerous parameters that can be set with *TranSim*, the most interesting parameters for our experiments are:

the protocol : TCP or UDP;

the TCP variant (if relevant): Reno [Jacobson, 1988], BIC [Xu et al., 2004], CUBIC [Rhee and Xu, 2005], HighSpeed [Floyd, 2003];

the rate limitation : none, PSP [Takano et al., 2005], HTB [17], TCP congestion window limitation, Iperf (for UDP traffic) [18].

Each of these parameters can be set separately for each node, or for groups of nodes.

In all of our experiments, the sources are nodes of the Lyon site, and Table 4.1 summarizes the specific characteristics of these source machines.

Rate limitation The last major aspect of the experimentation is the careful design of traffic generation. In real networks, flows are not the fluid ON/OFF flows: packets composing the flows are sent entirely, one after the other, at the wire bit-rate. Following up, a critical feature to consider in network-experiment design is the traffic-generation mechanism, especially the rate at which the packets are sent. For studies of 1 Gbps congestion with TCP sources, we want to let the TCP feedback mechanisms control the

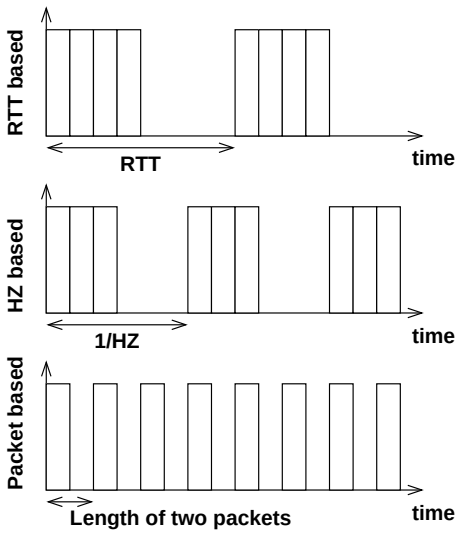

Figure 4.4: Different sending patterns. The upper figure shows a schematic view of an RTT-based rate limitation, the lower ones packet-level rate limitation, and in-between timer-based rate limitation.

emission rate, and we do not impose external source rate limitation. However, for the validation of Taqqu's theorem in loss-free conditions, that we perform in Section 5.1, we want to avoid congestion, and so we need to impose such an external source-rate limitation.

An important parameter is the aggregation level of traffic K, defined as the ratio between bottleneck capacity C and the access link's nominal capacity C_a [Srikant, 2007]. In xDSL context and more generally in the Internet, it is not rare to have K ranging over 1000, while in 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 (so as to mimic the natural rate limit due to the bandwidth of users' uplink in an Internet configuration) and to impose an aggregated throughput average lower than $C = 1$ Gbps, each source's rate has to be limited to at most 10 Mbps.

For the sake of scalability, only software end-hosts rate-limitation mechanisms are considered. We focus on GNU/Linux mechanisms as it is the most deployed operating system in our environment. End-host-based mechanisms can control the individual flows in a scalable and simple way [Soudan et al., 2007]. When considering packets of constant size, we can enforce data rates over a large period of time by imposing the inter-packets intervals. In end-host systems, four different beat times are controllable: a) userland timers, b) TCP self-clocking, i.e., RTT of the transfer's path, c) OS's kernel timers, and d) packet-level clocking. In our experiments we used 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, that is the RTT of the transfer's path (window size limitation).

The first two methods rely on Linux's traffic-shaping mechanism: with the `tc` utility [19], the qdisc associated with a network interface (the queue in which packets are put before being sent to the network card) is configured. The PSP (PSPacer) [Takano et al., 2005] qdisc spaces packets by inserting IEEE 802.3x PAUSE packets. These PAUSE pack-

ets 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 [17] 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. As the qdisc is set for the network interface, all the sources located on the same node will share the same bandwidth limit.

The third 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 $target\_throughput$. 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. For example, to limit the rate of a 1 Gbps source to 5 Mbps with full-size (1500 Bytes) Ethernet packets and a RTT of 12 ms, one has to fix the window size to 7.5 kBytes (corresponding to 5 full-size Ethernet packets). These values will be used in our validation of Taqqu's relation in loss-free conditions (Section 5.1).

Independently of this *artificial window limitation* that can be imposed to limit the source rate, the congestion window of a sender can also be limited by the receiver, through a maximal congestion window transmitted in the acknowledgment packets called *advertisement window* [Postel, 1981]. Even though the TCP header field corresponding to the advertisement window is only 16 bits, large values are possible with a scaling option [Jacobson et al., 1992] (the scaling factor in our configurations is 7). In all the configurations corresponding to our experiments, we verified (using web100 variables) that the advertisement window is never imposing a limitation below the capacity of the network card of the machines: 1 Gbps. This effect will then not be discussed in the rest of the manuscript.

Finally, for UDP transfer, the rate limitation is directly implemented in the *iperf* tool and relies on simple packet spacing (inter-packet times are determined so as to achieve the required bandwidth).

4.2.2 MetroFlux: high-speed packet capture

To complete the description of our experimental facility to study network traffic and QoS, we now describe our monitoring tool used to capture the traffic at a link up to 10 Gbps: *MetroFlux*.

Global architecture of the system

Metroflux is a programmable system for traffic analysis which currently operates on a 1 Gbps or 10 Gbps link without loss in most conditions. As described on Figure 4.5, this system is composed of hardware elements and software components. The *Metroflux* system integrates the GtrcNet-1 box [Kodama et al., 2003] or the GtrcNet-10 box [16] (for 10 Gbps links) and a storage server with a large amount of disk space. The system is able to capture the first 52 bytes (for GtrcNet-1) or 56 bytes (for GtrcNet-10) of every packet (*i.e.*, the header and a few payload bytes), add them a 1 μs-precision timestamp, and group them in a UDP packet that is sent to the capture server. This header aggregation reduces overhead, but also reduces the number of interrupts of the computer that receive the traffic to analyse, decreasing the local loss probability. The capture server runs a MAPI (Monitoring API [Polychronakis et al., 2004], developed by the LOBSTER project [21]) daemon with a

special GtrcNet driver that we developed, and is able to save the packet headers' stream into a pcap file for offline analysis (see Figure 4.6 for the detailed software architecture of the server). The pcap format has been chosen because it is also used by tcpdump and it saves packets with their capturing timestamp. The data-analysis server runs a packet header extracting program, based on the MAPI library which provides advanced functions to manipulate packet headers in pcap format. The maximal duration of a capture session depends on the packet size and the throughput of the link, as well as the server's storage capacity. It can encompass more than 60 hours for 1 Gbps traffic constituted of 1500 Bytes Ethernet packets, if the storage capacity is 1 TBytes. For 1 Gbps captures, our capture server is a computer with a quad-core processor running at 2.66 GHz, 4 GB of 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. For 10 Gbps captures, our capture server is a Bi-Dual-Core (Opteron) running at 3.0 GHz, with 8 GB of memory, 1 ethernet 10 Gbps port, 2 ethernet 1 Gbps ports, 146 GB disk for the system, and 15x300 GB SAS 15 Krpm for the captured data, connected via a hardware RAID 6 controller, for a total of 3.6 TB available for captured data (the partition for the captured data uses an XFS filesystem). The capture chain, including header extraction (GtrcNet), header grouping, header desencapsulation, and header storage in pcap format (capture server), has been extensively tested both for the 1 Gbps and 10 Gbps. The result is that even in the worst case (with 64 Bytes packets at the maximal rate), all the packet headers can be captured without loss with the 1 Gbps system. At 10 Gbps, our system can capture headers without loss in the case where 200 Bytes packets arrive at the maximal bandwidth of 10 Gbps. Note that this case is already very rare (most packets are 1500 Bytes), and that the system gives a warning for each lost packets, which never occurred in our experiments and production-links-monitoring campaigns.

The *Metroflux* system can be installed transparently within an experimental or a production network in two different ways (see Figure 4.7):

- incorporated in a 1 Gbps links (between source(s) and destination(s));
- plugged into a mirrored port of a switch.

If the *Metroflux* system is inserted in the data path, other GtrcNET functions such as latency emulation or aggregated-throughput on-line measurement can be activated. To fully appreciate the potential of our monitoring tool and its difference compared to other possible solutions based, for instance, on DAG cards, we develop in the next section the specificities of the GtrcNet device.

GtrcNet-10 characteristics for 10 Gbps links measurements

The FPGA-based hardware technology *GtrcNet* has been developed in AIST, Japan. Our part in the development of this device has been roughly limited to bug reporting, and we expose here the characteristics of the GtrcNet-10 only to give an overview of the functions available, and support our choice of using this device to integrate in *MetroFlux*, rather than other devices available, such as the DAG card, from Endace.

GtrcNet-10 [16] is a hardware layer-2 network equipment with 10 GbE ports (Gbps Ethernet ports). It is the successor of the GtrcNet-1 [Kodama et al., 2003] device, which has GbE ports. GtrcNet-10 provides many parametrable functions, such as traffic monitoring at a microsecond resolution, traffic shaping, and WAN latency emulation at a

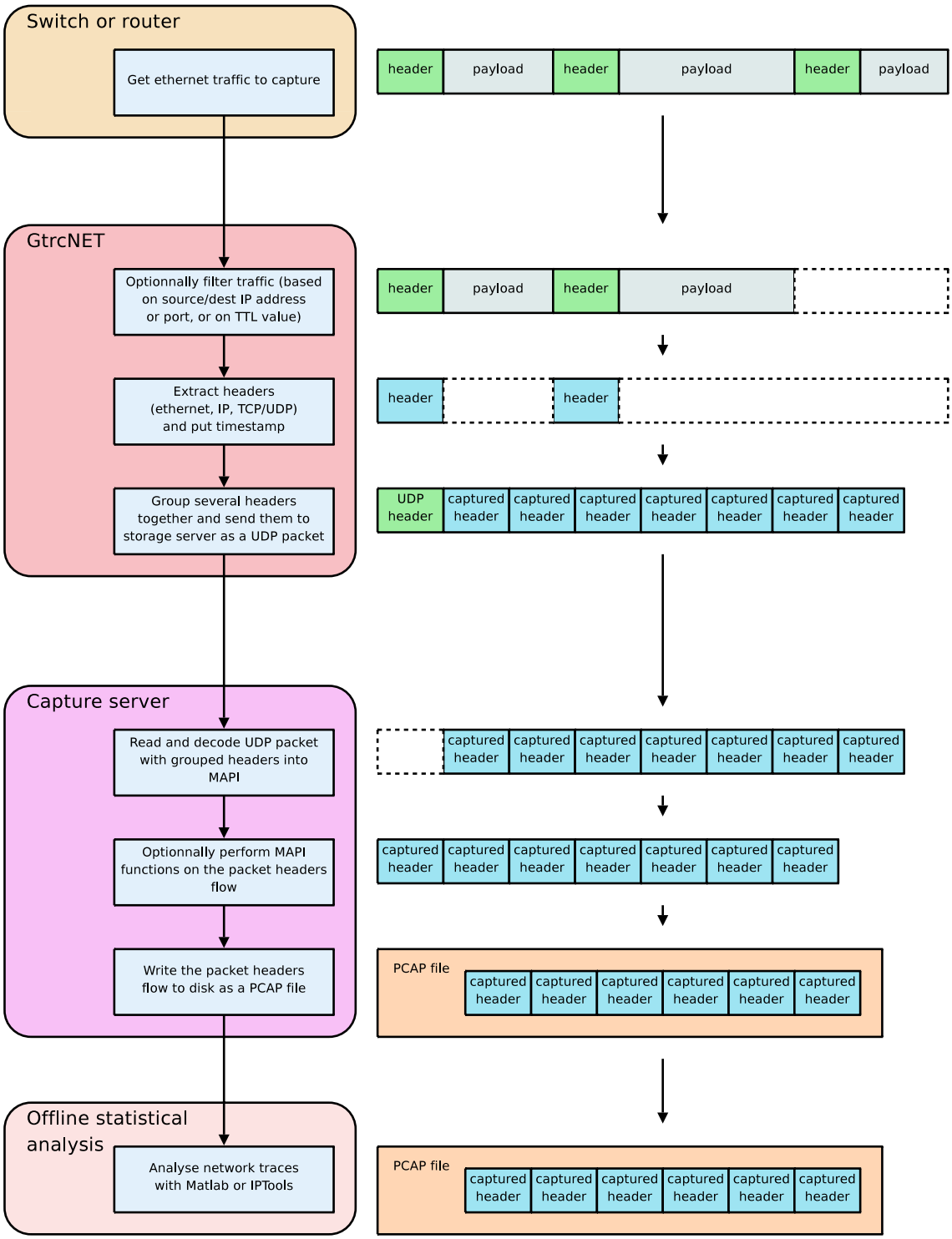

Figure 4.5: Capture and analysis of the traffic with *Metroflux*

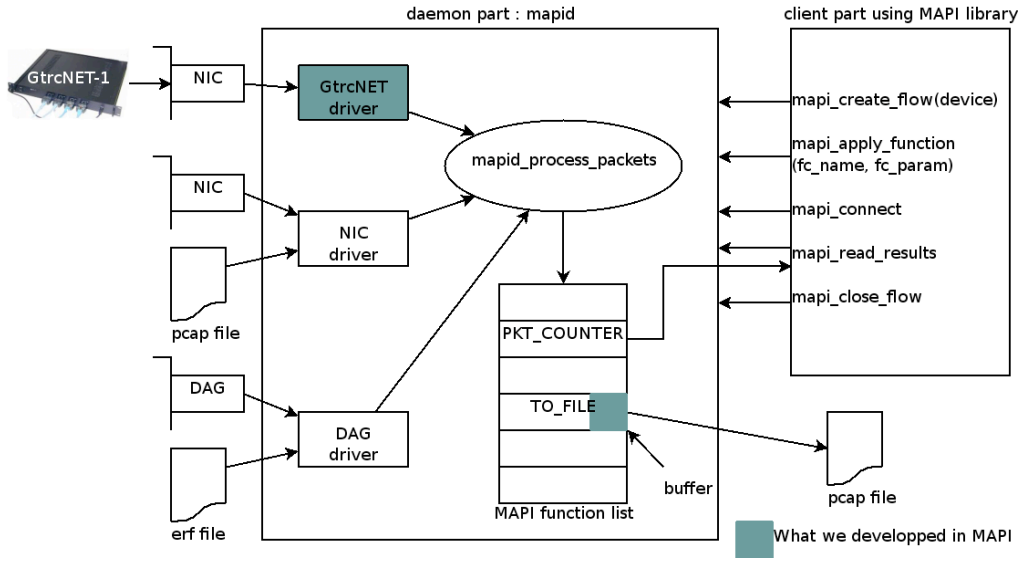

Figure 4.6: Software architecture of the server

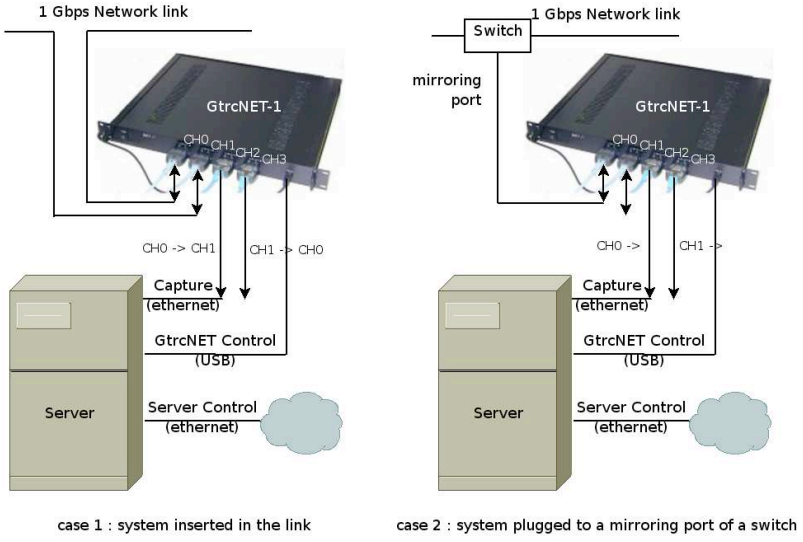

Figure 4.7: Two different ways of installing *Metroflux* on a link

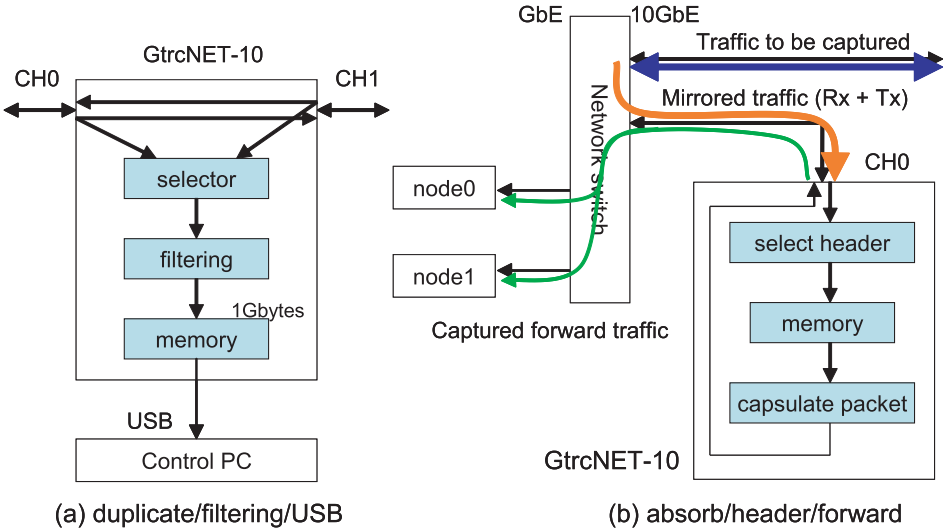

Figure 4.8: Example of packet capture on GtrcNet-10

10 Gbps wire speed. A remote computer can set several parameters, such as interval of traffic-bandwidth monitoring, target bandwidth of traffic shaping, and delay of network emulation. It gets results of bandwidth monitoring. In the *Metroflux* system, GtrcNet-10 is used and configured to capture headers of packets. GtrcNet-10 consists of a large-scale Field Programmable Gate Array (FPGA), three 10 Gbps Ethernet XENPAK ports, and three blocks of 1 GB DDR-SDRAM. By programming the FPGA, one can add new functions and improve existing functions according to the user's requirements.

Let us now detail the packet-capture functions of GtrcNet-10 with two examples. Figure 4.8 (a) shows a first example of the usage of GtrcNet-10 for packet capture. The traffic to be captured is transferred between ports CH0 and CH1. GtrcNet-10 selects an input port, duplicates packets from the port, and captures them in memory. This function cannot capture bi-directional traffic. The traffic transfered between ports CH0 and CH1 is not affected by the capture. A filtering condition is defined so that packets are captured only if they satisfy the filtering conditions. A condition indicates a 16-bit field of the header, selects any bits by 16-bit mask, and compares it with the specified 16 bit value. One or two conditions can be specified, and the logical combination of two conditions (logical-and or logical-or) can also be specified. For example, one can capture only packets whose source port or destination port of TCP/IP header coincides with a specified value. One can also capture only packets that have VLAN tag, the VLAN ID being a specified value. The captured data are sent to a control PC via USB. Since the USB access speed of the current implementation is only 100 Kbytes/sec, the capture size is limited to the memory size, which is 1 Gbytes. If all the packets of a wire-rate traffic at a 10 GbE link are captured, only 800 ms of traffic can be captured.

To overcome this limitation, Figure 4.8 (b) shows another example of usage for packet capture on GtrcNet-10 (the one that is actually used in *MetroFlux*). The traffic to be captured is mirrored by a network switch, or duplicated by a "network tap", and the mirrored traffic is transferred to GtrcNet-10. GtrcNet-10 can capture selected header fields of the packets into a memory. The fields to be captured are in 16-byte-long unit, and any

field can be selected independently up to 128 bytes. Each captured packet header receives a 64-bit timestamp, whose format is based on RFC-1305 [Mills, 1992] and it constitutes the first 8 bytes of the captured header fields for each packet. Multiple captured headers are encapsulated into a UDP packet, and transferred from an output port. Figure 4.9 shows the format of the UDP packet, which includes ten 80 bytes of captured data. The first block is a scheduled time for transmission. This field is only used inside of GtrcNet-10, and is not transmitted. The next three blocks are Ethernet, IP, and UDP headers respectively. The fifth block is a header of the forward packet that includes captured header size (hd), number of captured headers (num), sequence number of packet (seq), and its transmit timestamp (txTimestamp). Captured data follow the headers. If the one-way wire-rate traffic of a 10 GbE link with 1500 bytes IP packet is captured using the format in Figure 4.9, the capture forward traffic rate is about 556 Mbps. The forward traffic is read by a receiving node. If the traffic is too large to be read by single node, the destination of the forward traffic can be distributed to multiple nodes (up to 16) in round-robin.

By mixing received and transmitted traffic in the mirrored traffic, bi-directional traffic can be captured in a port, but the mirrored traffic is limited to 10 Gbps. Since a GtrcNet-10 can capture mirrored packets from three ports simultaneously, bi-directional traffic of wire rate can be captured by mirroring received and transmitted traffic independently. Notice that some switches cannot receive packets from the mirroring port. Since GtrcNet-10 can transmit capture forward packet from any port, it solves the problem, but the solution requires two ports of GtrcNet-10 to capture the traffic of one port.

GtrcNet-10 can capture packets in any combination, such as duplicating packets or not, filtering packets or not, selecting header fields or not, accessing by USB or forwarding as packets, but capture forwarding is only for selected headers. The forwarding function also offers the possibility of sampling the packet stream (so as not to exceed the storage capacity in long-duration captures, for example).

Double synchronized capture on GtrcNet-1

To study a buffer's dynamic, we need to perform two synchronized captures at the input and output of the switch, as presented on Figure 4.3. Several GtrcNet devices can be synchronized using GPS. However, using GPS necessitate local arrangements that are not always possible.

Thanks to the four GbE ports of the GtrcNet-1 (CH0-3), it is possible to simultaneously perform two captures on the same GtrcNet-1 device. The traffic for the first capture is received at port CH0 and headers are forwarded through port CH2, and the traffic for the second capture is received at port CH1 and headers are forwarded through port CH3. In addition to guaranteeing perfect synchronization of the two captures (all the timestamps are fixed by the same clock), this has the advantage that only one GtrcNet device is needed to perform two captures. With the GtrcNet-10 device, such a simple solution is not possible because it has only three 10 GbE ports. Other solutions are possible, that we are currently investigating.

4.2.3 Packet data processing

To go from the packet-level traces acquired with *MetroFlux* to the averaged and flow-level statistics of interest, we use a series of tools that we now present. Basically, we are

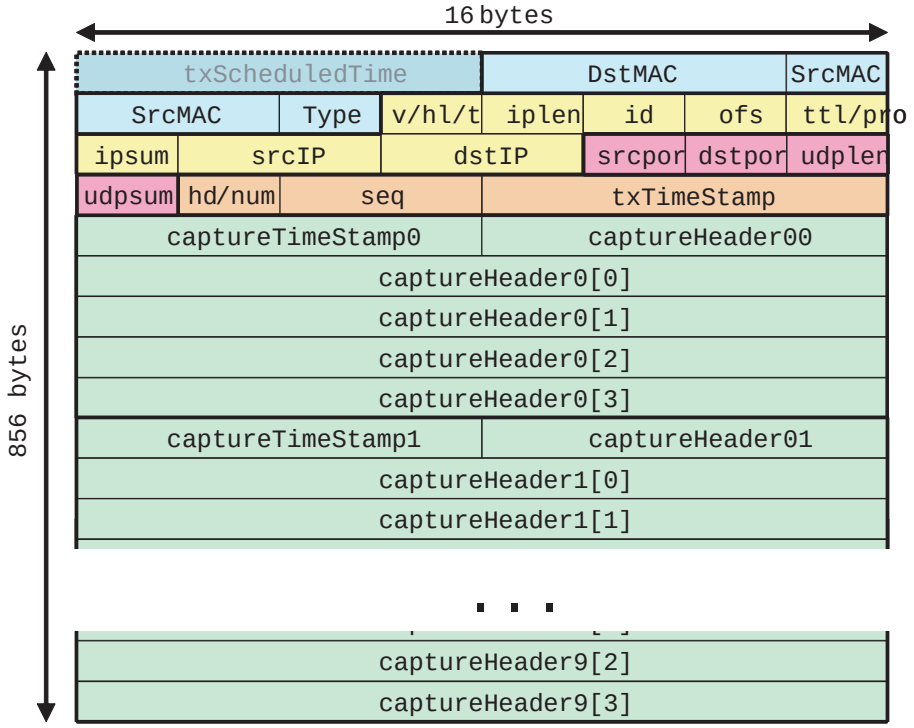

Figure 4.9: Packet format of capture forward

interested in recovering from the packet-level trace:

- the averaged bandwidth in consecutive time windows of arbitrary size Δ;
- the flow-level characteristics.

Moreover, extraction of subtraces for conditioned studies based on parameter filtering (e.g., traffic from/to a list of IPs, traffic on given ports, traffic using a specific protocol) can be required.

We use a series of tools over the captured IP traffic traces. A first step is to handle the captured IP traffic traces; secondly, we reconstruct the flows from the packets, and calculate the averaged bandwidth.

IP traffic traces, saved in standard pcap format by the capture device, are first processed by `ipsumdump`, a program developped at UCLA [20], 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 the needed information from our traces: timestamps, source and destination IPs, port numbers, protocol, payload size, TCP flags, and sequence numbers. The informations are condensed into a binary file that is easier to parse, and which does not depend on specific capture hardware anymore.

Secondly, we use a collection of tools named *IPTools* working on the `ipsumdump` binary format directly, which performs a variety of useful data operations on the traces. The *IPTools* software has been developed at the Physics laboratory of *ENS Lyon* and is registered at the APP[1] [Dewaele et al., 2007]. Of relevant interest here are the operations already

[1] Association de Protection des Programmes (French association for software protection)

mentioned: computation of the averaged-traffic time series; extraction of traffic subtraces for conditioned study, based on parameter filtering (e.g., traffic from/to a list of IPs, traffic on given ports, traffic using a specific protocol); 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 distribution heavy-tailedness. It is necessary to recompose each flow from the intertwined packet streams 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. This constraint faces the arduous issue of dynamic table updating.

In IPTools, flows are classically defined as a set of packets that share the same 5-tuple comprising: source and destination IPs, source and destination ports, and protocol. However, because there is a finite number of ports, it is possible for two different flows to share the same 5-tuple, 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 than `timeout` . Any subsequent packet with the same 5-tuple 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 flow. 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-tuple), and the FIN or RCT flag to close the flow, rendering `timeout` useless. 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 the currently open flows, using hash functions to speed up the access. Since TCP sequence numbers and payload size for TCP packets are captured, it is possible to search for dropped or re-emitted packets during the flow 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.

Other functionalities and data operations have been implemented. We only cite one here, namely, the extraction of traffic subtraces for conditioned study, based on parameter filtering (e.g., traffic from/to a list of IPs, traffic on given ports, traffic using a specific protocol).

IPTools also allows different kind of statistic computations on the data, such as time aggregation (including multi-scale time aggregation). The flows (or packets) can be aggregated at different timescales in a single pass, directly from the traffic data produced by ipsumdump. These operations can also be performed simultaneously on several subtraces of the traffic, in order to compare the statistical behavior of different parts of the data, or different kind of traffics.

Various other features have been implemented in IPTools, that we do not present here, because we do not use them in the rest of our work.

4.3 Real-traffic traces

To complement studies based on controlled experiments as presented in the previous section, we resort to real-traffic traces in two different University contexts. We now briefly present these traces.

In2p3 trace. The first trace, which is used in Section 5.3, is acquired at the output link of the in2p3 research center (Lyon, France) with our versatile monitoring tool *MetroFlux*. We "plugged" *Metroflux* at the output of the in2p3 to capture the traffic from the VLAN corresponding to RENATER web traffic. This VLAN is encapsulated in the 10 Gbps output link of in2p3 and we capture it with the *GtrcNet-10*. We captured this traffic during more than one day between January 18^{th} and January 19^{th}, 2009, and we observed a mean throughput slightly larger than 100 Mbps during working hours. This traffic mainly corresponds to web traffic, but has the particularity of including a larger number of elephants than usual traffic. Since the in2p3 is a nuclear-physics research center, these large transfers are likely to correspond to the transfer of experiment results from experiment centers like CERN.

ENS Lyon trace. The second trace, which is used in Chapter 6, comes from a different university environment: the 1 Gbps output link of the campus site of *École Normale Supérieure de Lyon* is monitored. The bidirectional traffic (input and output traffic) going through this link is optically splitted and captured with an appropriate DAG card [14] (see [Scherrer et al., 2007] for a more complete description of the system). The trace we consider in Chapter 6 was acquired on March 4^{th}, 2007, between 4:30pm and 5:30pm. It mainly consists of HTTP traffic from ENS students and researchers, achieving a mean throughput of 87.7 Mbps. This trace was acquired within the METROSEC project and is courtesy of the Physics lab. of ENS Lyon.

4.4 Numerical simulations

While analysis involving TCP traffic necessitates controlled experiments performed on a real network to comprehensively include all TCP parameters, and especially the feedback introduced; performing such experiments on a real network is less critical for UDP traffic. Indeed, the source emission in UDP is simply constant and independent of the rest of the system. Even though performing experiments in UDP with our metrology platform is possible, we also developed an ensemble of matlab routines to simulate traffic traces and queueing systems. Such numerical simulations can complement experiments performed on a real network and usually allows for more flexibility in the scenario design.

Our simulation tools consists in two major blocks: aggregate-traffic simulation and queueing simulation.

Our aggregate-traffic simulator can generate aggregate traffic from an arbitrary number of ON/OFF sources. The traffic is simulated at packet level so as to give a packet-level trace, basically equivalent to the traces that we acquire with *MetroFlux*. The simulator is very flexible and allows us to vary many parameters, the most important being:

- the flow-size- and OFF-times-periods distributions;
- each source's emission rate;
- the packet distribution within a flow.

Basically, any distribution that we are able to generate with matlab can be used for the flow size and OFF times. This includes the distributions of interest in our work, i.e., the exponential and heavy-tailed distributions, but also some distributions more complex to generate, e.g., the stable distribution, that are not directly available in ns2 for example.

Within each flow, the packets can be distributed in several ways. They can be regularly spaced, so as to mimic the UDP protocol's behavior (the inter-packet time is then determined according to the source's emission rate). They can also be placed according to a renewal process, again, where the inter-arrival time can follow any distribution available in matlab, provided that it has a mean. This mean inter-arrival time is then determined according to the source's emission rate. They can finally be distributed by bursts, so as to mimic the TCP protocol. This last way to place packets within a flow can be useful to analyze small-scale properties of TCP traffic caused by its burst mechanism, but we did not implement any feedback model. Indeed, as soon as it comes to studies involving TCP, we prefer to use controlled experiments on our metrology platform, rather than relying on a matlab program which would take into account the temporal dynamic of real systems. In our simulation design, a great attention has been devoted to the control of memory usage, so that we are able to generate very long traces, with intermediate storage on the hard disk, even on machines with limited memory. For each packet of the trace, we keep several quantities: a timestamp, the packets' size, an emission source ID and a flow ID number. The flow ID number allows us to perform simple flow-level analysis, particularly useful in our study implying sampling (Chapter 6). We also developed simple tools to recover from these packet-level traces the quantities of interest in our analysis, e.g., the averaged bandwidth in consecutive time windows of arbitrary size.

For QoS studies, we developed a discrete queue simulator adapted to our traffic simulator which simulates the dynamic of a single buffer fed by the output trace of our traffic simulator. For each packet, the queue simulator computes the delay if there is room available in the buffer, and signifies a loss otherwise. The size and the capacity of the buffer can be set to arbitrary values. To complement this discrete queue simulator, we also developed a continuous queue simulator, which takes as an input the traffic averaged in consecutive time windows of size Δ. The output is the loss rate and the buffer content sampled at Δ. We verified that, as long as Δ is smaller than the buffer scale (*i.e.*, time to feed or empty the buffer at a capacity C), both the continuous and discrete simulators give the same results. As we mentioned in Chapter 1, this is not surprising because the buffer has the effect of absorbing variations at a scale smaller than its characteristic scale. This advantage of the continuous time simulator is that it is much faster. The drawback is that individual packets cannot be tracked, for example to study the delay of packets pertaining to a given flow. We use these queue simulators in our study of the QoS in UDP, in Section 7.1.

Conclusion

The platform that we have presented in this chapter fulfills all of our requirements to study network traffic and QoS in realistic conditions: a large number of fully controlled machines, high-speed dedicated links and high-performance monitoring tools. Throughout the rest of our work, we will extensively use this unique platform to perform controlled experiments, and sometimes complement the controlled experiments by the real traces or simulations presented in this chapter. Table 4.2 summarizes the experiments and traces used in the rest of the manuscript.

Chap.	Sec.	Exp./Simul./Trace	Monitoring	Description
5	5.1	**Expe. validation of Taqqu's relation in the loss-free case**		
		Controlled exp.	*MetroFlux* 1 Gbps	Topology: 1 (Fig 4.2) Other site: Rennes Nb sources: 100 Protocol: TCP Rate limitation: 5 Mbps
	5.2.1	**Expe. validation of Taqqu's relation in the lossy-link case**		
		Controlled exp.	*MetroFlux* 1 Gbps	Topology: 2 (Fig 4.3) Other site: Rennes 10 Gbps cross tr.: constant UDP 1 Gbps monitored tr.: Nb sources: 50* ON/OFF protocol: TCP rate lim.: none
	5.2.2	**Expe. validation of Taqqu's relation in the congestion case**		
		Controlled exp.	*MetroFlux* 1 Gbps	Topology: 2 (Fig 4.3) Other site: Rennes 10 Gbps cross tr.: none 1 Gbps monitored tr.: Nb sources: 56* ON/OFF protocol: TCP rate lim.: none
	5.3	**Beyond Taqqu's theorem: correlation between flow size and rate**		
		in2p3 trace	*MetroFlux* 10 Gbps	see Section 4.3
6		**Estimation of α from packet-sampled traffic**		
		Simulations	–	Nb sources: 100 ON/OFF rate limitation: 10 Mbps regularly spaces pkts
		ENS trace	DAG 1 Gbps	see Section 4.3
7	7.1	**Analysis of QoS: the UDP case**		
		Simulations	–	Nb sources: 50 ON/OFF rate limitation: 38 Mbps regularly spaces pkts
	7.2	**Analysis of QoS: the TCP case**		
		Controlled exp.	*MetroFlux* 1 Gbps	Topology: 2 (Fig 4.3) Other site: Nancy 10 Gbps cross tr.: none 1 Gbps monitored tr.: Nb sources: 45 ON/OFF protocol: TCP rate lim.: none
8		**New scaling laws in TCP traffic**		
		Controlled exp.	*MetroFlux* 1 Gbps	Topology: 2 (Fig 4.3) Other site: Rennes 10 Gbps cross tr.: none 1 Gbps monitored tr.: 1–2 long-lived TCP sources and 1 constant UDP source or 32 ON/OFF UDP sources

Table 4.2: Summary of the traces and experiments used in the manuscript. Numbers of sources denoted by * use 2 sources by node.

Chapter 5

Large-scale self-similarity in aggregate network traffic: Taqqu's theorem and beyond

The results of Section 5.1 have been published in [1] [6] [8], the results of Section 5.2.1 are in [11], and the results of Section 5.3 are in [7].

Aggregate network traffic characterization.

- To what extend is Taqqu's relation valid in real network conditions? What are the scales involved? Does the protocol play a role? In which situation (loss-free, lossy link, congestion)?
- What happens if flow-duration and flow-size distributions have different tail indices? Which of these tail indices, if any, governs the long-range dependence of the aggregate traffic?

Introduction

In this chapter, we tackle the question of aggregate network traffic characterization, focusing on the large-scale scaling properties.

A major breakthrough in network traffic analysis and modeling has been the discovery of the self-similar nature of aggregate time series at large time scales [Leland et al., 1993, Paxson and Floyd, 1994]. Following up, the ON/OFF model proposed in [Willinger et al., 1995, Taqqu et al., 1997b] posits the heavy-tailed nature of the ON and/or OFF period as a plausible explanation for the origin of the self-similar property and clarifies the relation between the tail exponent and the Hurst parameter (see Section 3.1.1 of Chapter 3):

$$H = \frac{3 - \alpha_{\mathrm{H}}}{2}, \tag{5.1}$$

where $\alpha_{\mathrm{H}} = \min(\alpha_{\mathrm{ON}}, \alpha_{\mathrm{OFF}}, 2)$, α_{ON} (resp. α_{OFF}) being the tail index of the ON periods (resp. OFF periods) distribution. We call this relation (the same as relation 3.1 of Chapter 3) "Taqqu's relation".

Notwithstanding the mathematical soundness of this heavy tail/LRD model, the simplifications used to ensure its tractability (fluid source emission in particular), as well as the asymptotic nature of the result, make necessary an experimental investigation of the

conditions of its validity in real network situations. While some validations of Taqqu's relation (equation (5.1)) have been attempted (see Section 3.1.1 of Chapter 3), based either on real Internet traces or on simulators, none of them were combining all of the well-suited ingredients:

- a large-scale real experimental platform,
- a fully controllable environment,
- state-of-the-art estimators whose utilization is appropriately adapted to the experimental conditions;

and the precise conditions underlying validity of Taqqu's relation are still unclear. To answer this question, we proceed by steps: we first impose loss-free conditions (Section 5.1) and investigate the range of scales in which Taqqu's relation holds. We also elucidate the role of the protocol in this situation. Then we move to more realistic conditions, implying losses of two types. In Section 5.2.1, we impose losses after the aggregation point to emulate a lossy link, and in Section 5.2.2, we study the case of congestion losses arising at the aggregation point (which is then the bottleneck).

Another simplification of the ON/OFF model is that is assumes a constant emission rate, identical for each flow. This implies that the tail indices of the ON-time and flow-size distributions (α_{ON} and α_{SI}) are identical. Other models, mainly based on the infinite source Poisson model propose to randomly draw the rate of each flow, but they draw the rate independently of the size (or the duration) of the flows, so that these tail indices α_{ON} and α_{SI} are still identical (see Section 3.1.1 of Chapter 3). Such an assumption can fail in real traffic, yielding the question of the prediction of the large-scale self similarity in that situation where relation (5.1) cannot be used. In Section 5.3, we propose an extension of existing models, which accounts for the correlation between flow rates and durations (responsible of the different tail indices α_{ON} and α_{SI}). Our results in term of self-similarity index prediction are confronted to a real traffic trace exhibiting sharply different values of α_{ON} and α_{SI}.

5.1 Experimental validation of Taqqu theorem in the loss-free case

In the next two sections, we use the the potential and the facilities offered by the large-scale experimental platform presented in Section 4.2 to empirically investigate the scope of applicability of Taqqu's relation (equation (5.1)). Under controlled experimental conditions, we first prescribe the flow-size distribution to different tail indices and compare the measured self-similar exponents with their corresponding theoretical predictions. Our goal is not to check the goodness-of-fit between traffic data and ON/OFF models, but rather to investigate to what extent imposing such models on the sources of a real network facility produces LRD, as predicted by Taqqu's Theorem.

We start in this section with the advisedly *elementary* and yet *realistic* traffic pattern, where congestion is voluntarily avoided using rate-limitation mechanisms.

	description
TCP variant	Bic, with SACK
Topology	Butterfly (Figure 4.2)
Bottleneck	1 Gbps
Destination site	Rennes
RTT	12 ms
Sources nb.	100
Source rate	5 Mbps
Exp. duration	8 hours
Flows nb.	5.10^6
Aggregation time	$\Delta = 100$ μs
Flow timeout	`timeout` $= 100$ ms

Table 5.1: Fixed experimental global parameters.

5.1.1 Experiments' description

The experiments performed in this section use the topology of Figure 4.2 (Section 4.2 of Chapter 4). Table 5.1 summarizes the specific parameters used in these experiments.

We impose an ON/OFF emission scheme to 100 sources, with heavy-tailed ON and OFF distributions P_{ON} and P_{OFF}, and i.i.d. ON and OFF periods. The emission of packets in each flow is controlled by one of the methods described in Section 4.2.1, each source rate being limited to 5 Mbps to avoid congestion at the 1 Gbps bottleneck, and to guarantee loss-free traffic. In this situation, flow duration is proportional to flow size, at least for sufficiently large flows for which transient behaviors at the establishment of the target rate has negligible effect. Flow-size distribution P_{SI} and ON-times distributions P_{ON} are then identical, up to some constant multiplicative factor. Note also that we consider in this section and the following, the *size* of the flows, defined as the number of packets in the flow. Since the packets have a constant size of 1500 Bytes except for the first and last packets which may be smaller, both quantities are proportional, provided that the flow is long enough.

Let us emphasize that even though we impose a rate control (at TCP-window or packet level) to avoid congestion, our TCP sources remain realistic. Indeed, they are ack-clocked, they are regulated by slow-start and congestion avoidance algorithms of a real GNU/Linux protocol implementation, and they are subject to real packet regulation. In average, each packet goes through two layer-3 equipments (IP routers), four layer 2-equipments (1 Gbps or 10 Gbps Ethernet switches), and three layer-1 equipments (Optical CrossConnect).

All experiments consist of one trace of 8-hour traffic generation, representing a total of approximately $n = 5.10^6$ flows. We use the tools presented in Section 4.2.3 of Chapter 4, to extract flow level information (flow sizes and OFF periods), and the aggregate traffic time series $X^{(\Delta)}(t)$ (see the parameters used in Table 5.1).

In order to clearly define the terms of application of 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 5.2.

Experiment A: This is the cornerstone experiment to check relation (5.1). Distribution of the ON periods are prescribed to Pareto laws with mean $\mu_{\text{ON}} = 0.24$ s (corresponding to a

	Proto	Rate lim	α_{ON}	α_{OFF}	μ_{SI}	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	$\Delta_{j_1}^{\mu_{\mathrm{SI}}}$
C	TCP	TCP	-	1.1 – 4	1000	$\widehat{H}$
D	TCP	TCP	1.1 – 4	-	100	$\widehat{h}_{\mathrm{loc}}$
	UDP	iperf				

Table 5.2: Summary of experimental conditions.

mean flow size of $\mu_{\mathrm{SI}} = 100$ packets). The experiment is performed ten times with different prescribed tail indices α_{ON}, varying from 1.1 to 4. OFF periods are kept exponentially distributed with mean $\mu_{\mathrm{OFF}} = \mu_{\mathrm{ON}}$. For each value of α_{ON}, an experimental point $(\widehat{\alpha}_{\mathrm{ON}}, \widehat{H})$ is empirically estimated. Moreover, to evaluate the possible influence of the protocol, and of the rate-limitation mechanism, the same series of experiments is reproduced with UDP (user-level packet pacing) and TCP (using the 3 different rate-limitation mechanisms PSP, HTB and TCP congestion window limitation). The exact same trials of random variables defining the flow sizes and the OFF times 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 $\mu_{\mathrm{ON}} = 0.24$ s and $\mu_{\mathrm{ON}} = 2.4$ s, corresponding to mean flow sizes $\mu_{\mathrm{SI}} = 100$ and $\mu_{\mathrm{SI}} = 1000$ packets, respectively. The objective here is to relate μ_{SI} to the lower scale bound $\Delta_{j_1} = 2^{j_1}\Delta$ defining a sensible regression range to estimate H (see Section 2.4.3 of Chapter 2).

Experiment C: The protocol (TCP), the throughput limitation mechanism (TCP window limitation) and the mean flow size ($\mu_{\mathrm{SI}} = 1000$) being fixed, we investigate now the role of the OFF-periods distribution on the self-similar exponent H. Distribution of the OFF periods are prescribed to Pareto laws with mean $\mu_{\mathrm{OFF}} = 2.4$ s. The experiment is repeated with different prescribed tail indices α_{OFF}, varying from 1.1 to 4. ON periods are kept exponentially distributed with mean $\mu_{\mathrm{ON}} = \mu_{\mathrm{OFF}}$. For each value of α_{OFF}, an experimental point $(\widehat{\alpha}_{\mathrm{OFF}}, \widehat{H})$ is empirically estimated.

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 that of long-range dependance phenomena. The variable parameter is the ON distribution tail index as in experiment A, whereas 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 at small scales.

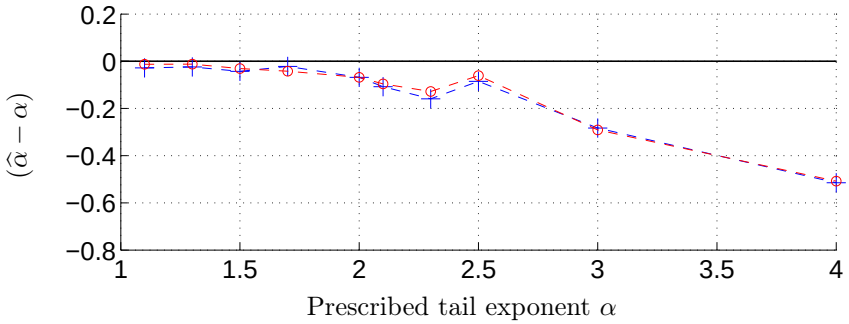

Figure 5.1: Difference between the prescribed HT index α and the actually estimated HT index $\widehat{\alpha}$ for the set of different values of α used in experiment series A (see Tabular. 5.2) for two different protocols: TCP (+) and UDP (∘).

5.1.2 Results and discussion

Verifying Taqqu's relation

For every trace, we use the wavelet-based methodologies described in Chapter 2 for heavy-tail and self-similarity analyses. The estimated tail index $\hat{\alpha}$ (corresponding to either $\widehat{\alpha}_{\mathrm{ON}}$ or $\widehat{\alpha}_{\mathrm{OFF}}$, clear from the context of Tabular 5.2) results from the linear regression of equation (2.22) (Chapter 2, Section 2.4.2) applied to the flow-size sequence (or to the OFF times series), where a sixth-order derivative of a Gaussian wavelet is systematically used. The self-similarity index $\widehat{H}$ is estimated from the LD plots of the aggregate traffic time series $X^{(\Delta)}$ (Chapter 2, Section 2.4.3), using a standard Daubechies wavelet with 3 vanishing moments [Mallat, 1999, Daubechies, 1992].

Tail-index estimates Proceeding with experiment A, for different values of the tail index of the flow-size distribution, Figure 5.1 displays the differences between the prescribed value α and the actually estimated value $\widehat{\alpha}$. The two experimental curves, corresponding to TCP and 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. 5.1 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, principally when the mean of the distribution is kept constant. Indeed, as α increases, large flows become more rare, and the number of observed elephants during the constant duration (8 hours) of the experiments decreases, progressively deviating from a statistically relevant sample. A second explanation lies in the finite sample size issue: as the distribution mean is fixed, the maximum observed value decreases and reduces the domain where the distribution asymptotically behaves as a power law. This observation is fully consistent with arguments developed in [Roughan et al., 1999]. In our numerical studies, we will then systematically compare the estimated Hurst parameter to $\widehat{\alpha}$ rather than to the prescribed α.

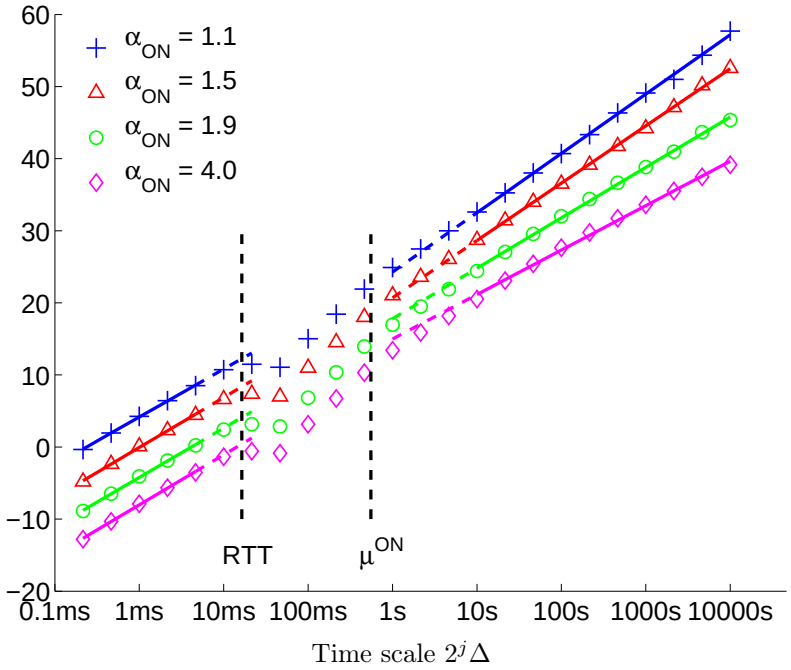

Figure 5.2: Wavelet log-diagrams of aggregate traffic (aggregation interval $\Delta = 100$ μs). Log-diagrams correspond to 4 time series obtained under similar experimental conditions (experiment series A) with the protocol TCP, with 4 different values of α_{ON}: 1.1 (+), 1.5 ($\triangle$), 1.9 ($\circ$) and 4.0 ($\Diamond$). For the sake of readability, curves were vertically shifted to avoid overlapping.

Gaussianity Before verifying the presence of long-range dependence in the aggregate traffic time series, we first need to verify the normal assumption (bear in mind that the increment process of a fBm is a stationary and Gaussian process). Thus, for each trace described in Tab. 5.2, the Kurtosis index[1] of the aggregate traffic time series distribution was computed at several different aggregation intervals. As the Kurtosis index was always found to lie between 3.0 and 3.1, for aggregation intervals larger than $\Delta = 10$ ms, we conclude that the aggregate traffic time series is a reasonably Gaussian process beyond this limit.

Without claiming to be a rigorous proof, this normality verification indicates that the number of used sources (100) is sufficient to meet the asymptotic conditions of an infinite number of users.

LD-description Figure 5.2 shows typical LDs of aggregate traffic time series, obtained under similar conditions (experiment series A of Tabular 5.2, TCP protocol, TCP window limitation), for 4 different values of α. Such plots enable a generic phenomenological description of LDs: 3 different ranges of scales can be visually identified, whose bounds do not seem to drastically vary with α:

Coarse scales: In the coarse scale domain, a clear scaling behavior is systematically ob-

[1] The Kurtosis index of a R.V. is defined as the ratio of its fourth-order moment over its square variance, and takes value 3 in the normal case.

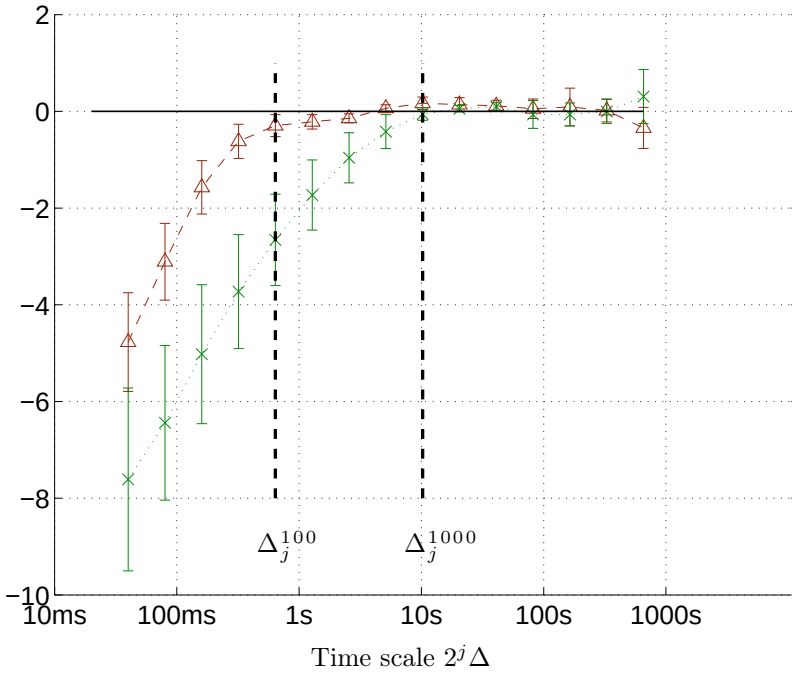

Figure 5.3: Averaged normalized log-diagrams for two different mean sizes μ_{SI} of the transmitted flows: ($\times$) $\mu_{\mathrm{SI}} = 1000$ packets – ($\triangle$) $\mu_{\mathrm{SI}} = 100$ packets.

served. Since Taqqu's Theorem (equation (5.1)) relates heavy tails and self-similarity in the asymptotic limit of coarse scales, the corresponding scaling exponent, denoted $\widehat{H}$, is a candidate to match that involved in relation (5.1).

Fine scales: At fine scales, another clear scaling behavior is also observed. However, the corresponding 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 noticeable scaling behavior.

In Figure 5.2, vertical lines materialize the two transition 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.

Coarse-scales domain's lower bound It is alluded in [Hohn et al., 2003] 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 experiment series with two different values of the mean flow duration (series B of Tab. 5.2). For each case, all the LDs corresponding to the different values of α are computed. To emphasize the impact of the mean flow duration, we substracted to each LD, the asymptotic linear trend, obtained by linear regression between a scale Δ_{j_1}, clearly above the *knee* position, and the maximum available scale $\Delta_{j_{\max}}$. Figure 5.3 shows, both for $\mu_{\mathrm{SI}} = 100$ and $\mu_{\mathrm{SI}} = 1000$ the mean and standard deviation over all normalized LDs. Each graph clearly exhibits a slope break: at scale $\Delta_j^{100} = 0.64$ s when $\mu_{\mathrm{SI}} = 100$ and at scale $\Delta_j^{1000} = 10.28$ s when $\mu_{\mathrm{SI}} =$

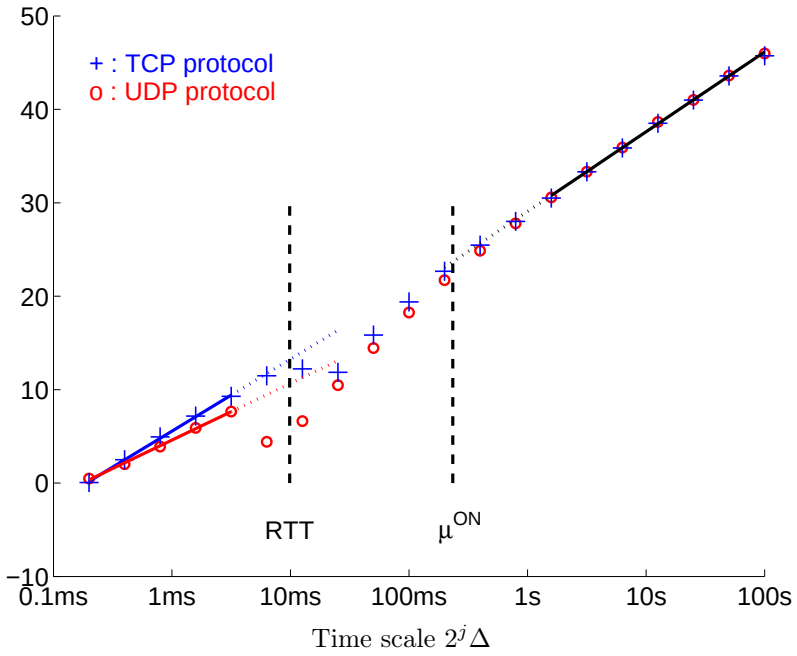

Figure 5.4: Wavelet log-diagrams of aggregate traffic (aggregation interval $\Delta = 100$ μs). Log-diagrams correspond to two time series obtained under similar experimental conditions, for $\alpha_{\mathrm{ON}} = 1.5$, with two different protocols: TCP (+) and UDP ($\circ$).

1000. Although for $\mu_{\mathrm{SI}} = 100$, the knee effect slightly smoothes out, the linear behavior observed for $\mu_{\mathrm{SI}} = 1000$ 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: $\Delta_j^{100} \simeq \mu_{\mathrm{ON}}^{100}(0.24\text{ s})$ and $\Delta_j^{1000} \simeq \mu_{\mathrm{ON}}^{1000}(2.4\text{ s})$. 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 renamed: the *flow scales*, or the *file scales*.

Note that the value μ_{ON} for the coarse-scales domain's lower bound is a rough bound. It has been recently shown in [Abry et al., 2009] that this bound can undergo small variations when α varies: the larger α, the larger bound.

Protocol, rate limitation and coarse scales We now focus on the coarse-scales domain. Figure 5.4 shows the LDs obtained with two different protocols: TCP and UDP (for $\alpha = 1.5$). It evidences the central feature that both LDs are undistinguishable in the coarse-scale domain. We conclude that, when source-rate limitation precludes congestion, the protocol has no impact on the coarse scale self-similarity.

Similarly, Figure 5.5 shows typical LDs ($\alpha = 1.5$, TCP) obtained with three different rate-limitation mechanisms: PSP, HTB and TCP window limitation. 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 influence on the scaling behavior at coarse scales.

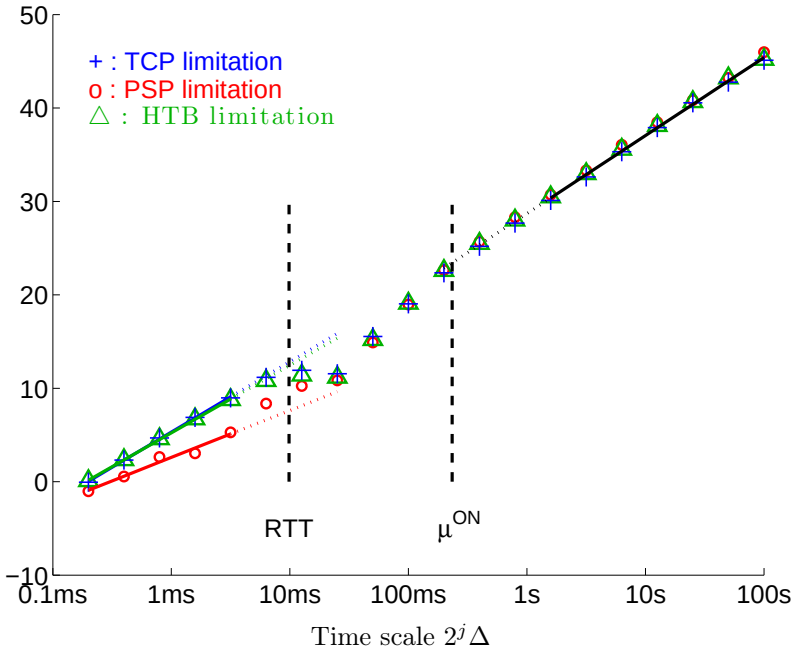

Figure 5.5: Wavelet log-diagrams of aggregate traffic (aggregation interval $\Delta = 100$ μs). Log-diagrams correspond to 3 time series obtained under similar experimental conditions, for $\alpha_{\text{ON}} = 1.5$, with three different rate limitation mechanisms: TCP (+), PSP (∘) and HTB (△).

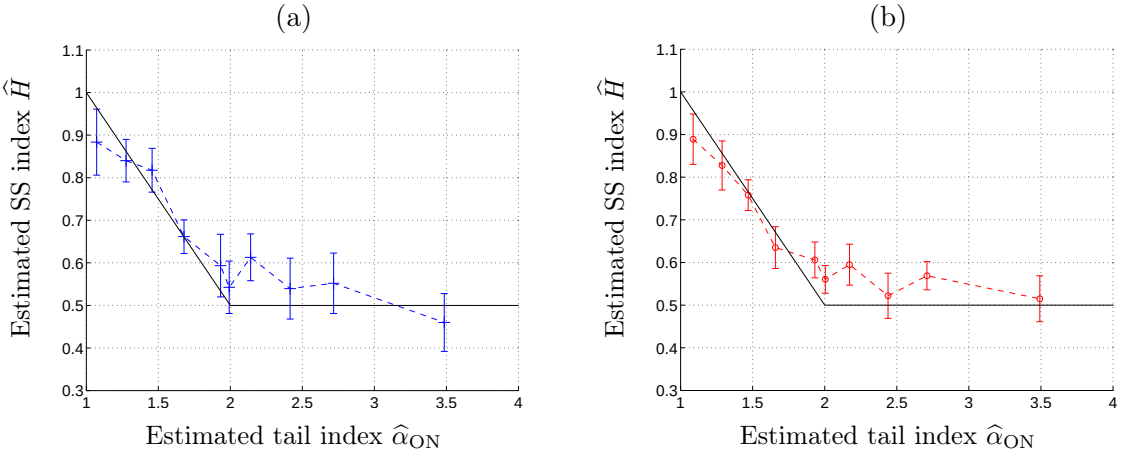

Figure 5.6: Estimated Self-Similar index $\widehat{H}$ of the aggregate traffic (aggregation interval $\Delta = 100$ μs) versus estimated tail index $\widehat{\alpha}_{\text{ON}}$ of the corresponding flow-size distribution. Solid plots represent the theoretical model of relation (5.1), dashed plots correspond to experimental results: (a) with the TCP protocol; (b) with the UDP protocol.

H versus α_{ON} Practically, to perform an empirical validation of relation (5.1), 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 found 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 self-similarity exponent estimates.

Figure 5.6 plots the estimates of coarse-scale scaling exponents against those of the tail indices. Confidence intervals for $\widehat{H}$ displayed on the graphs are supplied by the estimation procedure detailed in Section 2.4.3, under valid normal assumption. Such estimations are conducted independently for TCP and UDP protocols. For both protocols, estimations show a satisfactory agreement with Taqqu's theorem prediction (equation (5.1)). 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 deviation below the theoretical relation for α close to 1 has been significantly reduced when compared to similar analysis results reported in the literature (cf. *e.g.*, [Park et al., 1996]).

On the opposite, the origins of the difference between experimental and theoretical curves for α around 2, are clearly identified. As mentioned earlier, while the mean flow value is kept constant, the maximum flow size observed over a fixed period statistically decreases. This amplitude-range shrinkage mechanically imposes the LD scaling interval that conveys LRD to reduce as well[2], and thus the estimation of H to be overvalued. Another interpretation based on a bias-variance trade-off argument, and supported by matlab simulations, is also given in [Abry et al., 2009]. Notice though, that when α increases, this effect balances with the natural de-correlation ($H = 0.5$) holding beyond the maximum flow-size scale, and $\widehat{H}$ stabilizes to 0.5, which also coincides with the theoretical value predicted by Taqqu's theorem (equation (5.1)).

This improved agreement (as compared to [Park et al., 1996]) results from the combination of a number of factors, mostly: robust statistical estimators and long-duration controlled traces allowing us to really handle the asymptotic coarse-scale nature of Taqqu's theorem. Additionally, our analyses confirm that the TCP and UDP protocols do not impact this relation, at least under loss-free conditions corresponding to our experimental setup. This is in clear agreement with the findings reported in [Figueiredo et al., 2005], 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), and that congestion is properly avoided, the only cause for self-similarity is the heavy tail in the flow-size distribution.

[2]Indeed, at scales beyond the maximal flow duration, the aggregate throughput is necessarily uncorrelated.

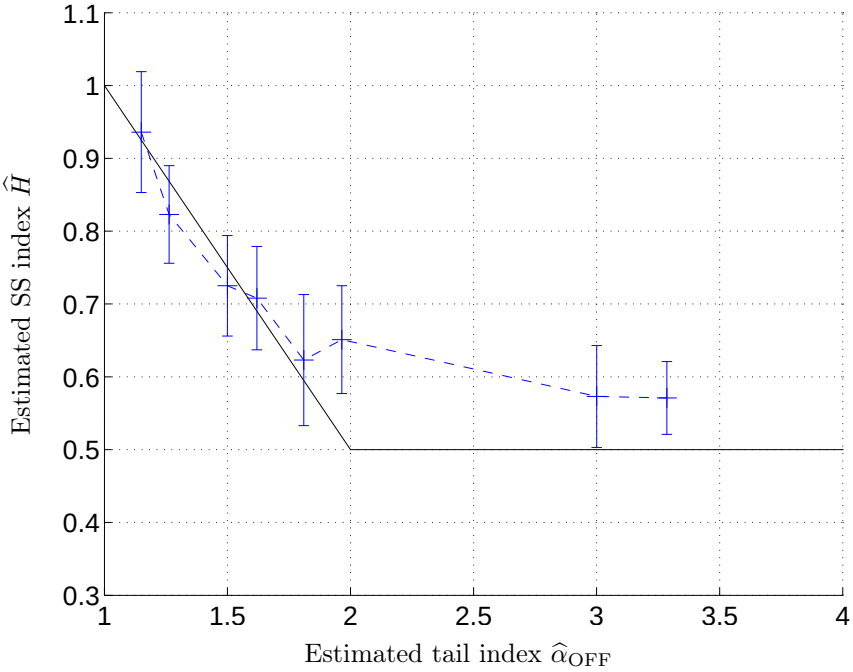

Figure 5.7: Estimated Self-Similar index $\widehat{H}$ of the aggregate traffic (aggregation interval $\Delta = 100$ μs, TCP) versus estimated tail index $\widehat{\alpha}_{\mathrm{OFF}}$ of the corresponding OFF-periods distribution. Solid plots represent the theoretical model of relation (5.1), dashed plots correspond to experimental results.

OFF periods Under experimental conditions detailed in Tabular 5.2, Figure 5.7 displays the estimated coarse-scale self-similarity exponents against those of the OFF-periods tail indices. Since we already demonstrated that the protocol has no influence on the scaling behavior at coarse scales, the current experiments were only performed with the TCP protocol.

In contrast to similar results reported in the literature (cf. *e.g.*, Figure 5 (right) of [Park et al., 1996], where the estimated value of H is less than 0.7, even for $\alpha_{\mathrm{OFF}} = 1.05$), our results show a very good match with the theoretical relation of equation (5.1). Reasons for this theoretical accord certainly has the same origins as the ones evoked in the previous paragraph (*i.e.*, robust estimators, long-duration stationary traces and controlled configurations), but possibly also in the fact that we scrupulously avoided congestion, and consequently statistical alteration of the OFF periods. To the best of our knowledge, this other part of Taqqu's Theorem had never been satisfactorily addressed.

Finally, let us notice that confidence intervals displayed in Figure 5.7 are significantly larger than the ones of Figure 5.6. This is due to the difficulty of imposing short OFF intervals which led us here to increase the mean durations $\mu_{\mathrm{OFF}} = \mu_{\mathrm{ON}} = 2.4$ s (instead of 0.24 s). Consequently, according to the interpretation of Figure 5.3, the coarse-scale regression range proportionally narrows.

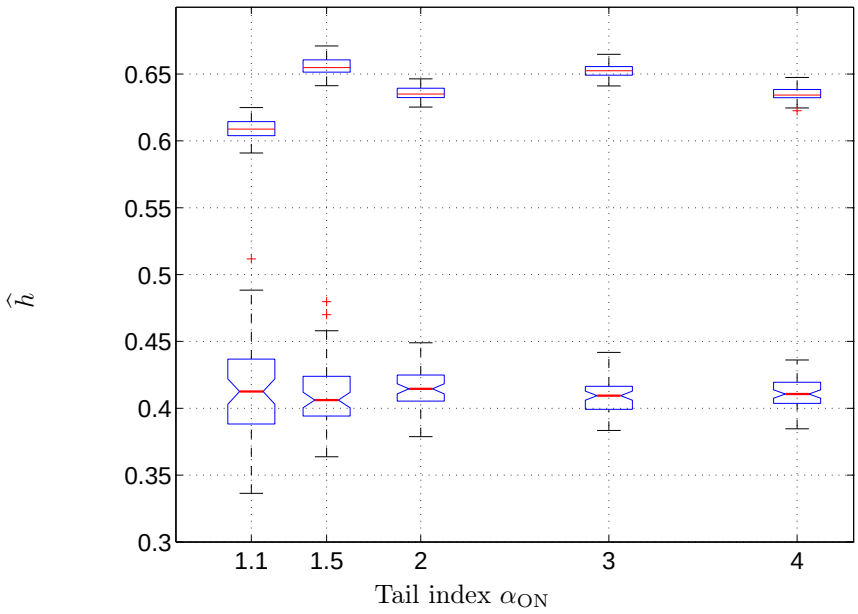

Figure 5.8: Fine-scale scaling exponent h estimates on aggregate traffic time series ($\Delta = 100\mu s$). For different values of the tail index α_{ON} governing the flow-size distributions, h is estimated by linear regression of Log-diagrams (see Fig. 5.4) over the scale range $[0.2 - 5]$ms. Notched box-plots correspond to the UDP protocol, regular box-plots to the TCP protocol.

Further analyses of the LD

We now turn to the medium and fine scales and study the influence of protocols and rate-limitation mechanisms.

Medium scales

While the mean flow duration gives an upper bound for the medium-scale domain, Figure 5.2 shows that RTT (12 ms) corresponds 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. 5.4 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. 5.5 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 Mbps 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 μs, which is much lower than RTT. Accordingly, the rate limitation does not impact the traffic at RTT scales.

Fine scales

Without entering the debate about potential multifractality of network traffic at small

scales[3], we discuss now the second-order self-similarity observed at fine time scales.

TCP and UDP impact on fine-scales scaling. Figure 5.4 shows a good scaling behavior at fine scales, with different scaling indices for UDP and TCP.

To analyze the fine-scales scaling exponent in more details, every 8-hour trace corresponding to a particular value of α (see experimental conditions of Experiment A in Table 5.2) is chopped into 66 short-length segments of duration $T = 100$ s each. The resulting time series are then analyzed independently and a fine-scaling exponent h estimated. Based on these 66 values of $\widehat{h}$, box-plots are displayed on Figure 5.8 for each theoretical value of α. The values for TCP remain roughly constant around $h \simeq 0.63$. Likewise for UDP, h does not seem to depend on α, but stands around 0.4, a significantly smaller value than that for TCP.

Smaller than RTT, these fine scales correspond to the *packet-scales*. Clearly then, 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 maintain the rate limitation (5 Mbps). Therefore, 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 of 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's impact on fine-scales scaling. Figure 5.5 shows that the scaling index at fine scales 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, and the local regularity is then identical. On the contrary, when using PSP, packets are sent individually, in the same way as with UDP, and the small-scale scaling index with UDP and PSP are very close, lower than the ones observed with TCP and HTB. These results are fully coherent with the arguments developed in [Jiang and Dovrolis, 2005].

5.1.3 Conclusion

In this section, we experimentally demonstrated the validity of Taqqu's theorem (equation (5.1)) in loss-free situations. Based on realistic and long-duration stationary traces, we obtained a significantly better match than previous results existing in the literature, especially in the case where the OFF periods are heavy-tailed. As it is elegantly tackled in [Guo et al., 2001] and [Figueiredo et al., 2002], we believe that a precise statistical characterization of idle times is an important factor of traffic modeling. Finally, following up the controversial discussion about the relationship between transport protocols and self-similarity (see, *e.g.*, discussion in [Figueiredo et al., 2005]) our observations confirm that in loss-free situations, protocols and rate-limitation mechanisms do not impact the observed long-range dependence. Our analyses show that this is so, because the ranges of scales

[3] Performing a multifractal analysis of our traces of loss-free traffic would not be of great interest. Indeed, the rate-limitation mechanism prevents sources from congestion and this suppresses the role of the TCP feedback, usually invoked to explain multifractal properties of the traffic.

(segmented according to the *RTT* and to the mean flow duration) related to self-similarity are far coarser than those (fine and medium scale) associated to such mechanisms. Based on this supervised study in loss-free situations, the causes of a possible break down of Taqqu's relation in real-world traffic will have to be sought elsewhere and in particular, the protocol in itself should not be solely incriminated.

5.2 The effect of losses on LRD

We now use the loss-free study performed in previous section as a reference to investigate the validity of Taqqu's relation (equation (5.1)) in more realistic conditions implying losses. We distinguish two cases: the lossy link and the congestion losses.

5.2.1 The lossy-link case

We first treat the case of the "lossy link", where losses occur after the aggregation point. Strictly speaking, a lossy link would be a link where packets are lost at random, independently, *i.e.*, with Bernoulli losses. To emulate such a loss process, we provoked congestion after the aggregation point, as we now explain.

Experiments' description

For the experiments presented in this section, we use the topology of Figure 4.3 (see Section 4.2), with 50 "1 Gbps monitored TCP sources" (2 sources per node). The aggregate traffic from these sources shares the principal router of Lyon with constant UDP cross traffic from 20 machines (the "10 Gbps cross traffic"). This UDP cross traffic creates packet losses after the aggregation point of the TCP sources (at the principal router of Lyon). Indeed, the aggregate TCP traffic is always less than 1 Gbps, so that there is no loss at the aggregation point (the Lyon switch of Figure 4.3). Varying the sending rate of the UDP machines (x Mbps) allows us to vary the loss rate experienced by the TCP flows. The destination site is Rennes and the *RTT* experienced by the TCP flows can vary, due to the congestion at the Lyon router, between 12 ms and its maximal value of 50 ms. On the nodes running TCP flows, we collect *netstat* statistics to compute the loss rate. Table 5.3 summarizes the parameters of the experiments.

In all the experiments, the flow sizes are heavy-tailed distributed with tail index $\alpha = 1.5$ and the mean flow size is $\mu_{\mathrm{SI}} = 1000$ packets. The OFF times are exponentially distributed, with mean $\mu_{\mathrm{OFF}} = 0.6$ s. We use the same random sequences of flow sizes and OFF times in three experiments with different loss rates summarized in Table 5.4.

The first experiment is performed without UDP cross-traffic to avoid packet loss. In this experiment, we impose a source rate limitation using the TCP window limitation method, to avoid congestion at the Lyon switch. The limited rate is set to 20 Mbps for each source, which yields a mean ON time μ_{ON} equal to the mean OFF time. For the last two experiments (with losses), the TCP window is "not artificially limited", but is naturally limited by the TCP congestion-control mechanism. Table 5.4 shows the mean ON and OFF times measured from the traces, where the flows were reconstructed with an infinite `timeout` . We used an infinite `timeout` here as the reference because we are considering TCP traffic only (most of the flows can be separated with the SYN and FIN packets), but we also verified that a `timeout` of 1 s leads to the almost same flow sequence.

	description
TCP variant	Bic, with SACK
Topology	Butterfly (Figure 4.3)
Bottleneck	10 Gbps (after aggregation)
Destination site	Rennes
RTT	12–50 ms
TCP sources nb.	50
Fl. size distr.	heavy-tailed, $\alpha_{SI} = 1.5$
Mean flow size	$\mu_{SI} = 1000$
OFF times distr.	exponential
Mean OFF time	0.6 s
Exp. duration	2 hours
Flows nb.	10^6
Aggregation time	Δ=10 ms

Table 5.3: Fixed experimental global parameters.

	UDP rate (x)	loss rate	μ_{ON} (s)	μ_{OFF} (s)
$\alpha = 1.5$	–	no loss	0.53	0.76
	450 Mbps	0.7%	0.55	0.77
	475 Mbps	5%	6.14	0.70

Table 5.4: Experiment parameters: loss (estimated from netstat logs) and flow size and OFF time mean (estimated from the flows' reconstruction with an infinite `timeout`).

For the third experiment, although the imposed flow sizes' sequence is the same as for the first two experiments, the mean ON time is significantly increased, due to the relatively high loss rate.

Results and discussion

LD of the aggregate traffic We first investigate the LDs of the aggregate traffic time series $X^{(\Delta)}(t)$ for the three experiments, displayed on Figure 5.9. The mean ON time for each experiment has been materialized for it has been identified (see Section 5.1.2) as the lower bound of the coarse-scale domain of interest for the long-range dependence of relation (5.1). We now focus on these coarse scales.

While a clear scaling behavior is observed for the three experiments, the corresponding Hurst parameter is much smaller for a high loss rate of 5%, than for small loss rates. Table 5.5 displays the estimated Hurst parameter of the three experiments and confirms this observation: while Taqqu's prediction ($H = 0.75$, see equation (5.1)) is very well verified for small loss rates (up to 0.7%), a high loss rate (5%) annihilates the long-range dependence of the aggregate traffic (H very close to 0.5). This is not due to the increase of the mean ON time because the Hurst exponent is estimated in a significant scale range lying far beyond this lower bound in each experiment.

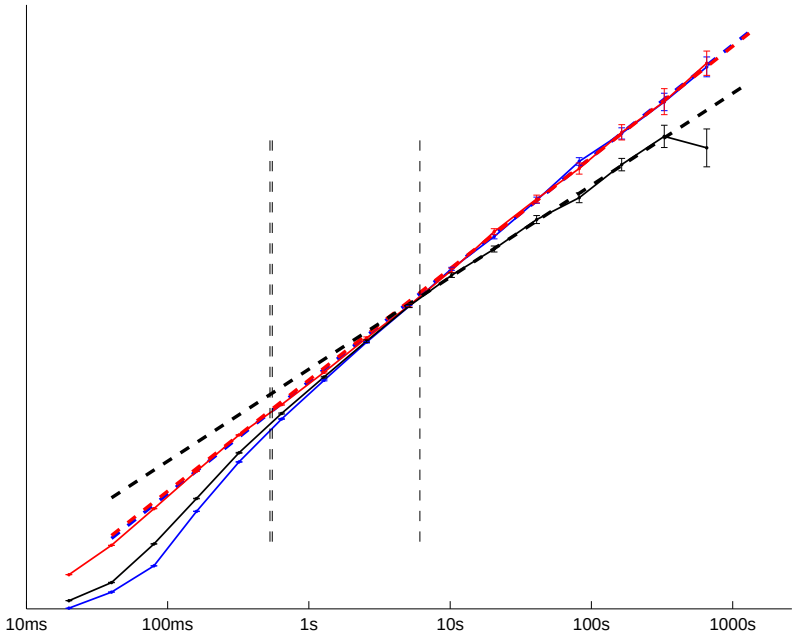

Figure 5.9: Log-Diagrams of the aggregate traffic time series $X^{(\Delta)}(t)$for the three experiments with $\alpha = 1.5$: (blue) no loss – (red) loss rate 0.7% – (black) loss rate 5%. The vertical lines materializes the mean ON times as shown in Table 5.4.

	loss rate	$\widehat{H}$
$\alpha = 1.5$	no loss	0.76 ± 0.05
	0.7%	0.75 ± 0.07
	5%	0.53 ± 0.06

Table 5.5: Estimated LRD indices

Interpretation We now propose an interpretation of this disappearance of the long-range dependence for high loss rates, which is fully coherent with Taqqu's theorem. The key point of this interpretation lies in the flow reconstruction, and more particularly in the choice of the `timeout` .

Figure 5.10 displays the flow-size and OFF-time distributions observed in the three experiments after the flow reconstruction with two different values of the `timeout` : infinite `timeout` and `timeout` =100 ms. This latter value was chosen because it is smaller than the mean OFF time (about 0.6 s), but remains larger than the RTT. For this reason, it allows us to detect gaps inside a flow that are greater than 100 ms and split this flow into 2 different flows, but does not separate each TCP window in a single flow.

For the first two experiments (no loss and loss rate 0.7%), the distributions reconstructed with the two values of `timeout` are almost the same, showing that almost no flow experiences an internal gap of more than 100 ms. For a small loss rate of 0.7%, it shows that the TCP congestion window rarely falls down to zero, or at least, rarely stays at zero for more than 100 ms. Since the reconstructed flow sequences are almost identical with the two values of `timeout` , the OFF distributions in these two experiments are also very similar.

On the contrary, with a loss rate of 5%, the distributions reconstructed with the two values of `timeout` are very different: the flow-size distribution appears heavy-tailed with an infinite `timeout` , but not with a `timeout` of 100 ms (Figure 5.10 (bottom, left)). The strong modification of the flow-size distribution with a `timeout` of 100 ms shows that the flows experience internal gaps of duration larger than 100 ms. These gaps correspond to the intervals between two packet retransmissions during the *TCP exponential backoff periods* [Paxson and Allman, 2000]. Exponential backoff periods arise when no acknowledgment packet is received during an RTO period. This can in particular happen if all the packets of a congestion window are lost. Then, the congestion window is set to one packet which is sent, and the RTO is doubled (justifying the term *exponential* backoff). The durations of the gaps within flows thus take the values $2^k \times RTO_0$, $k = 0, 1, 2, \ldots$, where RTO_0 is almost equal to $4RTT$s, corresponding to about 200 ms with congestion. These characteristic values 200 ms, 400 ms and 800 ms indeed correspond to the peaks on the OFF-time distribution observed on Figure 5.10 (bottom, right) in addition to the imposed distribution with a `timeout` of 100 ms. Figure 5.11 shows that the flow-size distribution obtained with a `timeout` of 100 ms is exponential. For large original flows (when reconstructed with an infinite `timeout`), smaller flows reconstructed with a `timeout` of 100 ms mainly correspond to the packets emitted between two consecutive exponential backoff periods. Under Bernoulli loss assumption, their exponential distribution can then be interpreted using the fact that the exponential distribution is the only memoryless distribution. Indeed, assuming that a flow has achieved some stationary regime (for example with a mean throughput as in the square-root formula [Padhye et al., 1998]), then the probability to experiment a TCP timeout (no ACK received during an RTO) does not depend on the amount of data already transfered, so that the flow-size distribution is memoryless. Such an interpretation is likely to be valid only for the tail of the new distribution obtained with a `timeout` of 100 ms. This tail is however sufficient to predict the long-range dependence parameter via Taqqu's relation (equation (5.1)). As the internal gaps within flows are of the same order of magnitude as the mean OFF time, the flows reconstructed with an infinite `timeout` cannot be considered as single ON periods. Instead, flows reconstructed with a `timeout` of 100 ms must be considered to use the ON/OFF model and interpret

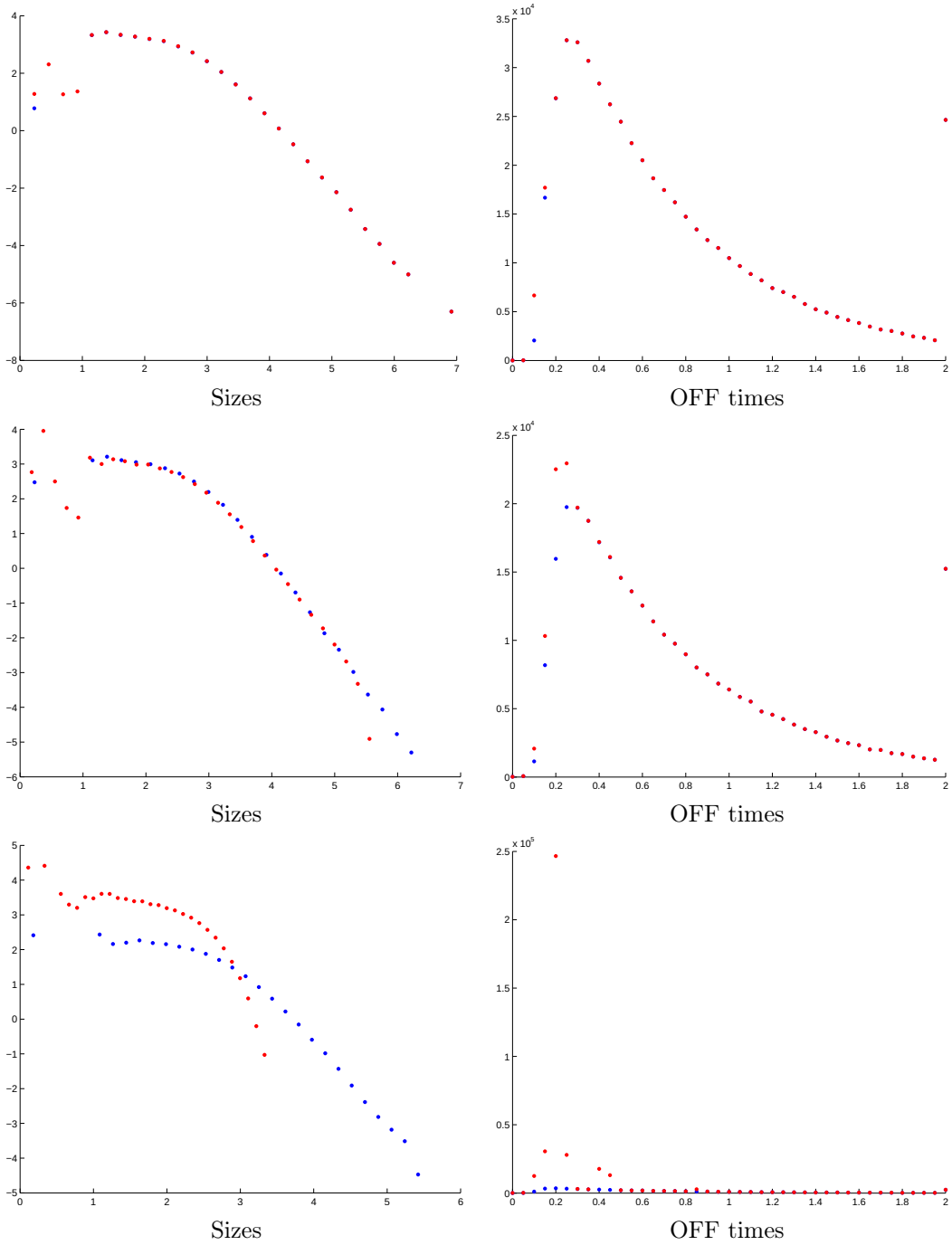

Figure 5.10: (left) Flow-size distribution in log-log – (right) OFF time distribution. (top) no loss – (middle) loss rate 0.7% – (bottom) loss rate 5%. Flows are reconstructed with: (red) a `timeout` of 100 ms – (blue) an infinite `timeout` .

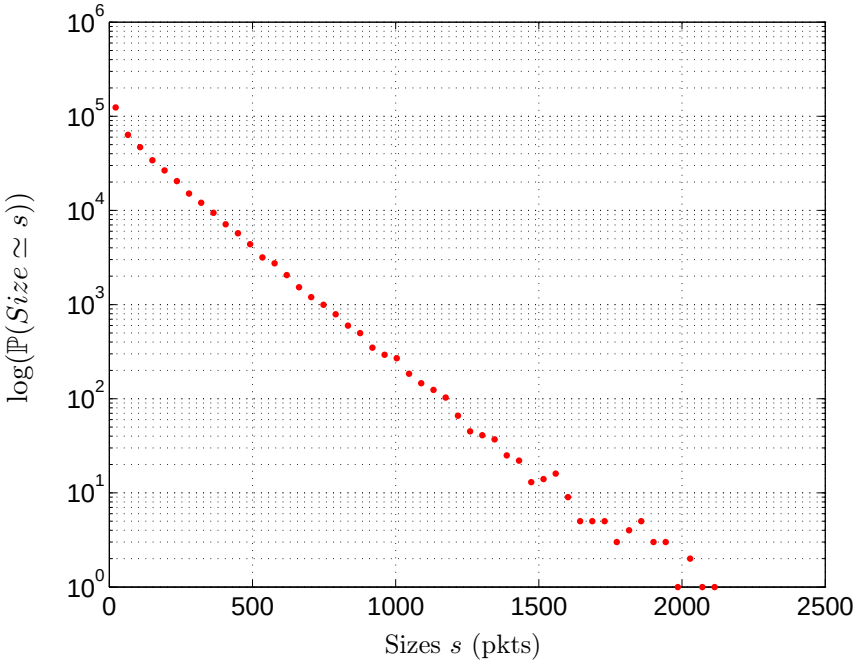

Figure 5.11: Flow-size distribution in semilog axes in the experiment with a loss rate of 5%, and where flows are reconstructed with a `timeout` of 100 ms (same distribution as in Figure 5.10 (bottom, left, red) with different axes to show the exponential distribution).

the LRD index using Taqqu's Theorem. The estimated Hurst parameter $H = 0.5$ when the loss rate is 5% is then in perfect agreement with the observed exponential distribution.

Conclusion

We showed in this section that in the case of a lossy link, high loss rates (5% in our experiments) annihilate the coarse-scale long-range dependence originating in heavy-tailed flow-size distributions. We showed that this results from a deep modification of the flow structure, due to the exponential backoff periods, splitting a single flow in several smaller flows, so that the flow distribution eventually becomes exponential. This study clarifies the notion of flow in the ON/OFF model (packets should be "regularly" arriving during a ON period with no internal gap of the order of magnitude of the mean OFF time), and gives practical guidelines for the choice of an adequate `timeout` for the flow reconstruction.

From a more general viewpoint, our result shows that under certain conditions, TCP mechanisms (the exponential backoff, here) are able to annihilate long-range dependence in the aggregate traffic. It might also be invoked to explain the relatively poor match between experimental curves and Taqqu's relation (equation (5.1)) observed in [Park et al., 1996], where the loss rate sometimes achieves 4%.

A more comprehensive set of experiments would be required to precisely quantify the critical loss rate marking the transition between the regimes where long-range dependence is observed or not, and to precise what it depends on (the mean OFF time, the flow-size distribution's tail index, etc.). Instead of using 10 Gbps cross traffic to emulate the lossy link, we will perform these experiments using the capability of the *GtrcNet* device to

	description
TCP variant	Reno, with SACK
Topology	Butterfly (Figure 4.3)
Bottleneck	1 Gbps
Destination site	Nancy
RTT	10–12 ms
Sources nb.	56
Fl. size distr.	heavy-tailed, $\alpha_{\text{SI}} = 1.1 - 3.0$
Mean flow size	$\mu_{\text{SI}} = 1000$
OFF times distr.	exponential
Mean OFF time	0.6 s
Exp. duration	4 hours
Aggregation time	$\Delta = 10$ ms
Flow timeout	`timeout` $= 100$ ms

Table 5.6: Fixed experimental global parameters.

emulate pure Bernoulli losses.

5.2.2 The case of congestion

In this section, we turn to the case of "congestion losses": the losses occur at the aggregation point (which is thus the bottleneck). This case is different from the lossy-link case because it introduces correlation between the sources. This correlation between the sources is not taken into account in Taqqu's ON/OFF model, and we again use an empirical approach based on controlled experiments, as we now expose.

Experiments' description

To perform this section's experiments, we use the topology displayed on Figure 4.3. As our goal is to create congestion at the 1 Gbps Lyon Switch, we do not impose 10 Gbps cross traffic. We use 56 "1 Gbps controlled sources" (with 2 sources per node), with no rate limitation, emitting to the destination site of Nancy. The minimal RTT is 10 ms, and the buffer size is 96 packets, susceptible to increase the RTT to no more than 12 ms (the maximal queueing delay is 1.2 ms). Table 5.6 summarizes these fixed parameters.

We perform 4 experiments, imposing an ON/OFF emission scheme to the sources, with a heavy-tailed flow-size distribution of variable tail index ($\alpha = 1.1, 1.5, 1.9$ and 3.0). The OFF-time distribution is always exponential, of mean $\mu_{\text{OFF}} = 0.48$ s. Since the source rate is not limited and is adapted by TCP mechanisms, cannot impose exactly the mean ON time. We impose a mean flow size of $\mu_{\text{SI}} = 1000$ pkts, which yields a mean ON time measured in the experiments roughly always equal to the mean OFF time.

In all of our experiments, the loss rate observed is between 0.4 and 0.5%.

Results and discussion

Figure 5.12 shows the aggregate traffic of all the 56 sources, averaged in consecutive time windows of variable size Δ for the four values of α. It clearly shows that for each experi-

ment, the aggregate traffic saturates at the bottleneck capacity (1 Gbps, corresponding to 8333 packets per 10 ms on the graphs), even when the traffic is averaged at large time scales ($\Delta = 10$ s). As a consequence, this traffic is non Gaussian (we always observed Kurtosis larger than 3.6 on the traffic shown on Figure 5.12), and has an asymmetrical marginal distribution. This saturation effect comes from TCP mechanisms (AIMD in particular), which try to keep the aggregate bandwidth constant, at the link capacity. Note that in our experiments, the pure truncation effect due to lost packets does not significantly modifies the traffic shape because the loss rate remains small (around 0.5%). Thus, the input and output traffic at the bottleneck are fairly similar. Ideally, the saturation effect would be characterized by the constraint that the sum of the TCP congestion windows at each instant are equal to the capacity of the link, which induces correlation between the sources at the RTT scale (TCP's characteristic time scale). In practice, non-instantaneous TCP reactivity permits smaller values of the sum of the congestion windows. Figure 5.12 shows that while the random ON/OFF structure introduces some variability in the aggregate traffic, the correlation between sources at RTT scale affects the aggregate traffic even at large time scales ($\Delta = 10$ s) where the saturation effect is still visible.

Since the total aggregate traffic is not Gaussian, even at large time scales, we do not try to measure long-range dependence on it, nor to validate Taqqu's relation (equation (5.1)). Indeed, even if some long-range dependence related to Taqqu's model was present, it would certainly be masked by the saturation effect at the link capacity. Instead, we consider the sub-traffic corresponding to half of the sources (*i.e.* 28 sources chosen at random within the 56 sources). Whereas for the ensemble of all the sources, the saturation effect constraints the non Gaussian marginal distribution of the aggregate traffic, the constraint is different for half of the sources: the aggregate traffic should equal the link capacity minus the aggregate traffic of the other half of the sources, which is also a random variable. Thus, the aggregate traffic corresponding to half of the sources has "one more degree of freedom", and it can be expected to be Gaussian despite the RTT-scale correlations between sources. Figure 5.13 shows the aggregate traffic of half of the sources, and exhibits no visual saturation effect. Moreover, we verified that for all the traffic shown on Figure 5.13, the Kurtosis always lies between 2.9 and 3.0. This good Gaussianity is the first step of Taqqu's model validation on the data.

To complete the investigation of Taqqu's model, Figure 5.14 shows the Hurst-parameters estimates against the estimates of the ON-time distribution tail indices. We verified that despite the flow rate variability due to TCP control mechanisms, the estimates of the flow size distribution tail index always coincides with that of the ON time distribution tail index. Moreover, the tail indices estimated from both halves of the traffic are also identical.

Although Figure 5.14 shows a clear decreasing tendency of the Hurst parameter, we do not observe a very good match between the experimental curve and the theoretical prediction (equation (5.1)), contrarily to the loss-free case (cf. Figure 5.6). Based on the results discussed in Section 5.1, we can eliminate a number of causes to explain this deviation. Indeed, since the same wavelet tools have been used for Hurst-parameter and tail-index estimations in Figure 5.6 and Figure 5.14, poor statistical estimations cannot explain the deviation of Figure 5.14. Similarly, long-duration stationary traces have been used in both cases: although the experiment time is only 4 hours in the congestion case instead of 8 hours in the loss-free case, it is unlikely to explain the deviation of Figure 5.14, *a fortiori* because the Hurst parameter has been estimated in a significant scale range ($[1.28 - 163.84]$ s) lying

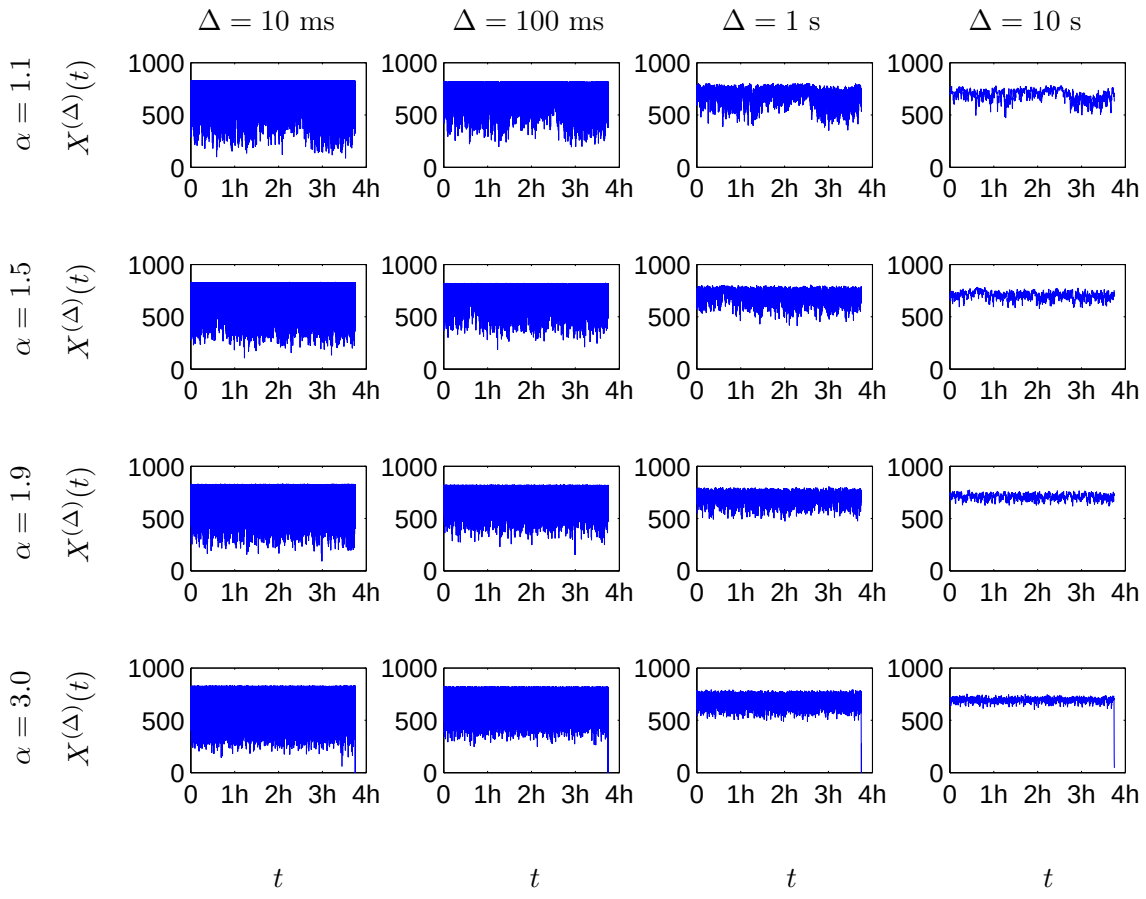

Figure 5.12: Aggregate traffic of all the 56 sources: Count process $X^{(\Delta)}(t)$ (in packets per 10 ms) versus time t, for different values of Δ and α. Note the saturation at 8333 pkts/10 ms, corresponding to the bottleneck capacity (1 Gbps).

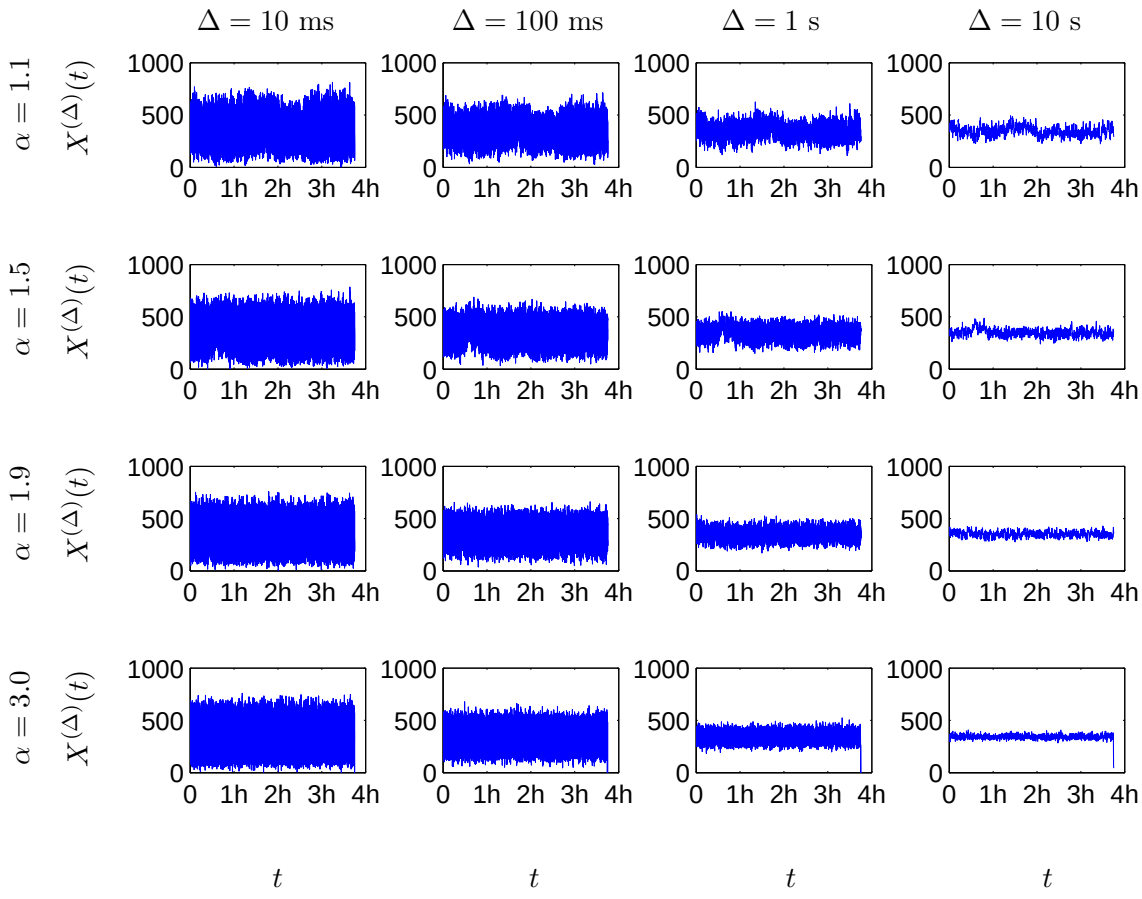

Figure 5.13: Aggregate traffic of 28 randomly chosen sources: Count process $X^{(\Delta)}(t)$ (in packets per 10 ms) versus time t, for different values of Δ and α.

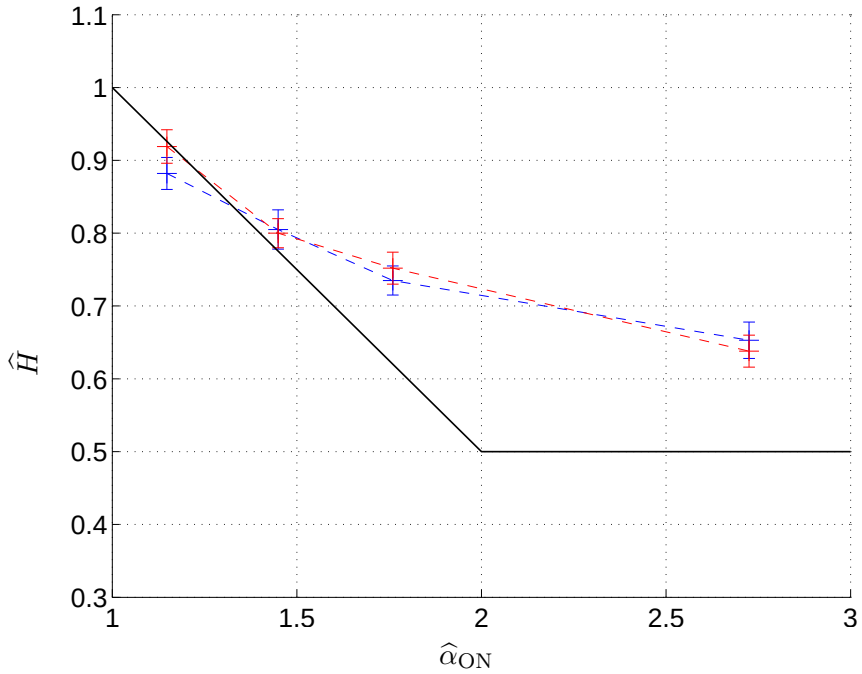

Figure 5.14: Estimated Self-Similar index $\widehat{H}$ of the aggregate traffic versus estimated tail index $\widehat{\alpha}$ of the corresponding flow size distribution. Solid plots represent the theoretical model of relation (5.1), dashed plots correspond to experimental results: (blue) 28 randomly chosen sources – (red) the other 28 sources. The scale range of estimation of H is $[1.28 - 163.84]$ s.

far beyond the mean ON time and exhibiting a good linear behavior. A possible explanation of the deviation between theoretical and empirical curves on Figure 5.14 could then lie in the correlation between sources, reminiscent from the saturation effect induced by TCP feedback mechanisms. Since we have seen that this correlation, although taking place at the RTT scale, is able to modify the total aggregate traffic even at large time scales, it is not surprising that for half of the traffic, this effect also holds, and that long-term correlations are then increased.

Conclusion

In this section, we considered the case of congestion losses, introducing correlation between the sources due to their interaction at the bottleneck. We showed that it introduces a saturation effect on the total aggregate traffic which is then non-Gaussian and unlikely to enter Taqqu's model framework. Instead, we focused on the aggregate traffic of half the sources and showed that it is Gaussian, and exhibits long-range dependence. However, we observed a significantly larger deviation with respect to Taqqu's relation (5.1) than in the loss-free case, and we proposed to interpret it by the RTT-scale correlations between the sources which can affect the traffic properties even at large time scales.

A more complete set of experiments is required to confirm and clarify our interpretation. In particular, half of the sources in our experiments here represent only 28 sources,

which might partly explain the deviation observed on Figure 5.14. Moreover, longer experiments should be performed in order to specify until which scale the RTT-scale correlation introduced by TCP has an effect. Indeed, our experiments proved that this correlation impacts the aggregate traffic up to a very large scale (of about 1 hour here), but we cannot eliminate the possibility that at even larger time scales, the effect of the effect of this correlation vanishes.

For a better understanding, a model taking into account the RTT-scale correlation is needed in order to get more general insights into the aggregate-traffic properties at each time scale. Such a model, to the best of our knowledge has not yet been proposed.

5.3 Beyond Taqqu's Theorem

In all the experiments presented in the last two sections, the tail indices of the flow-size distribution and the ON-time distribution was equal. In this final section, we treat the case where these indices are different, due to correlation between the flow rates and flow durations. This case does not fit in the framework of Taqqu's ON/OFF model, nor that of any other model proposed up to now (see Chapter 3, Section 3.1.1).

Our study is motivated by the observation of a real web traffic trace, which exhibit clearly different tail indices for the ON period and flow size distributions. We then propose a model accounting for the correlation between the flow rates and durations, which allows us to predict the Hurst parameter of the aggregate traffic (the equivalent of relation (5.1) of the ON/OFF model). We finally use our results to interpret the long-range dependence observed on our real traffic trace.

5.3.1 A real traffic trace: description

In this section, we use the real traffic trace corresponding to the web traffic of in2p3, Lyon, France (see a general description in Section 4.3 of Chapter 4). Although we captured the traffic during more than one day, we restrict our analysis to a short part of the trace to obtain stationary traffic, and use only the incoming traffic between 3pm and 3:30pm on January 18, 2009. The mean throughput in this period is 127.3 Mbps. Contrary to previous sections, the "size" of flows for this trace are in Bytes, and not in packets. To reconstruct the flow sequence, we use a `timeout` of 100 ms.

Figure 5.15 shows the flow-size and flow-duration distributions. While both distributions clearly appear to be heavy-tailed, they exhibit sharply different tail indices: $\alpha_{\mathrm{SI}} = 0.8544$ and $\alpha_{\mathrm{ON}} = 1.1994$ respectively. Several explanations could be posited to interpret why longer flows achieve higher rates, thus explaining these different tail indices. A might explanation might lie in transient effects at the beginning of each flow. Another explanation could lie in the variable locations of the downloaded files, if for example users tend to download large flows from closer locations (thus achieving higher rates more quickly because of smaller RTTs). The largest files might also be intentionally stored on servers with the highest capacities. These are only hypotheses and we do not elaborate further here on the origin of the different tail indices for the flow-size and flow-duration distributions, which is out of our scope.

Figure 5.16 displays the log diagram of the aggregate traffic bandwidth in time windows of size $\Delta = 10$ ms. We clearly observe a coarse-scale scaling behavior, with an estimated Hurst parameter of $\widehat{H} = 0.901 \pm 0.045$. The estimation is performed in the scale range

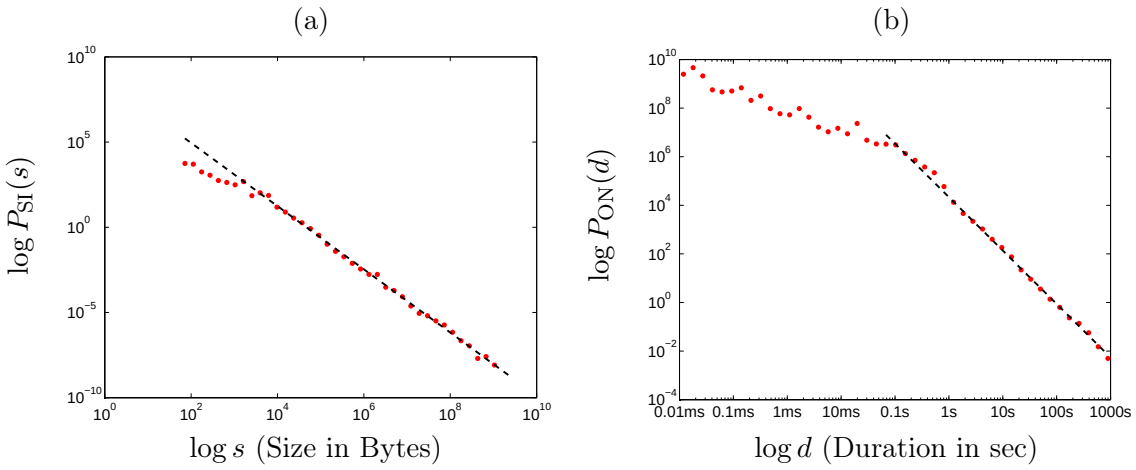

Figure 5.15: (a) Flow-size distribution – (b) Flow-duration distribution. The tail indices estimated with the wavelet method are respectively: $\alpha_{SI} = 0.8544$ and $\alpha_{ON} = 1.1994$. Straight lines materialize these slopes and have been vertically adjusted to the data. The mean ON time is 0.12 s.

$[0.64-40.96]$ s (lying far beyond the mean ON time of 0.12 s) but the same scaling behavior appears to extend to the scale 100 s. This Hurst parameter shows a good agreement with relation (5.1) of Taqqu's ON/OFF model, where $\alpha_H = \alpha_{ON}$ would be the tail index governing the long-range dependence. However, as already mentioned, the trace studied here is not well modeled by the ON/OFF model with constant rates, and drawing general conclusions from this case study would certainly be hazardous. Instead, we propose to develop a model taking into account the different tail indices of the flow-size and flow-duration distributions that originate from the flow rates and durations correlation.

5.3.2 A model accounting for the correlation between flow rates and durations

Definitions and notations

Our model is based on the infinite source Poisson setting: flows are arriving as a Poisson process. As we have seen in Chapter 3, such a model is not able to take into account the long-range dependence of the flow arrival process (should this be at the origin of the long-range dependence observed in the aggregate traffic). However, we observed on the real trace of in2p3 that the flow-arrival process has a coarse-scale parameter around 0.65, which is unlikely to explain the long-range dependence observed on the aggregate traffic bandwidth.

Figure 5.17 depicts our setting. The point process (T_i, D_i) (representing the arrival time and the duration of the flows) is assumed to be a planar (spatial) Poisson process of measure Λ [Cox and Isham, 1980, Kingman, 1993]. This means in particular that the sequence $(D_i)_{i\geq 0}$ is independent. The measure Λ determines the mean number of points in a part of the plane. For example, $\Lambda(C(t_1))$ is the mean number of points in the cone $C(t_1)$ (see Figure 5.17). In the case where the measure has a constant density (or intensity), the point process is called homogeneous. Here, the measure Λ contains the information of the

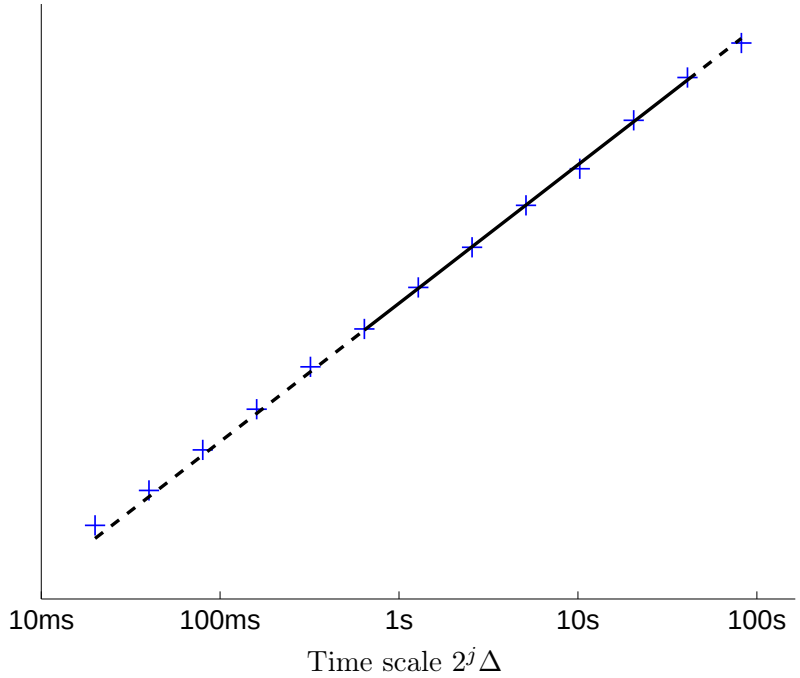

Figure 5.16: Log diagram of the aggregate traffic bandwidth in time windows of size $\Delta = 10$ ms. The estimated value of the Hurst parameter is $\widehat{H} = 0.901 \pm 0.045$ (scale range of estimation: $[0.64 - 40.96]$ s)

flow-duration distribution. To account for a heavy-tailed flow-size distribution, we use the following form for the measure Λ of an elementary square of size $(\mathrm{d}t, \mathrm{d}d)$ centered on (t, d):

$$\Lambda(\mathrm{d}t, \mathrm{d}d) = \frac{C \mathrm{d}t \mathrm{d}d}{d^{\alpha_{\mathrm{ON}}+1}}, \text{ where } C > 0, \text{ and } \alpha_{\mathrm{ON}} > 1. \tag{5.2}$$

Due to the dependence of Λ in the variable d, the point process is non-homogeous (the density of points is higher for smaller values of d). However, due to the independence of Λ in the variable t, the arrival time of a flow T_i is independent of its duration D_i. In our case where the x-axis represents the time, this specific form of the measure Λ ensures the stationarity of the resulting traffic. To avoid integrability problems for d around zero, we also set a minimal flow duration ε, under which the distribution is 0. This threshold will not play any role in our study because we concentrate on long-range dependence. A similar setting is used for example in [Barral and Mandelbrot, 2002], where the behavior of the measure around $d = 0$ has a great role because the authors focus on small-scale multifractality (note however, that in this paper the authors study a multiplicative process which differs from our additive process considered here).

Each flow emits data at a constant rate R_i during time D_i. The rates are random variables such that sequence $(R_i)_{i\geq 0}$ is independent. However, contrarily to previously proposed models (see in particular [D'Auria and Resnick, 2006, D'Auria and Resnick, 2008]), we do not assume that R_i is independent from D_i. Instead, we permit some correlation between the flow rates and durations that we will specify later in this section, when it becomes necessary to pursue the calculations. The flow size, that we denote S_i is then determined by $S_i = D_i R_i$.

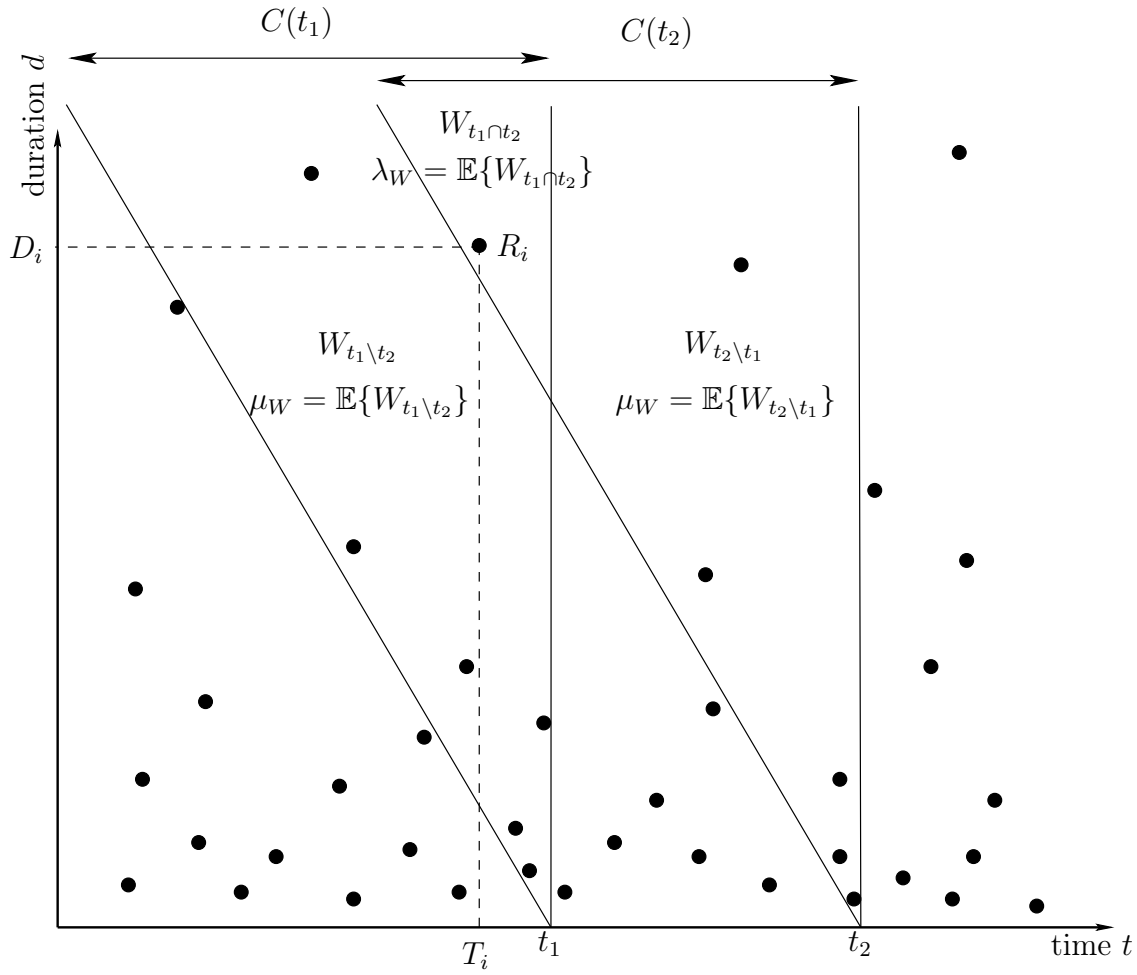

Figure 5.17: Setting of the model. Each point represent a flow. The x-coordinate represents the start time T_i and the y-coordinate the duration D_i. R_i is the rate of the flow. At time t_1, the active flows are those lying in cone $C(t_1)$ (the left border of the cones have a slope of -1).

We consider two time instants $t_1 < t_2$ and introduce the following notations (see Figure 5.17):

$$W_{t_1\backslash t_2} = \sum_{(T_i,D_i)\in C(t_1)\backslash C(t_2)} R_i, \tag{5.3}$$

$$W_{t_1\cap t_2} = \sum_{(T_i,D_i)\in C(t_1)\cap C(t_2)} R_i, \tag{5.4}$$

$$W_{t_1} = W_{t_1\backslash t_2} + W_{t_1\cap t_2} = \sum_{(T_i,D_i)\in C(t_1)} R_i. \tag{5.5}$$

The variables $W_{t_2\backslash t_1}$ and W_{t_2} are defined similarly. The random variable W_{t_1} is directly the instantaneous throughput at time t_1 (the sum of the rates of the flows active at time t_1). The random variables $W_{t_1\backslash t_2}$ and $W_{t_1\cap t_2}$ are intermediate random variables, useful for the calculations developed in the following. We can interpret these random variables as follows: $W_{t_1\backslash t_2}$ corresponds to the traffic of the flows that are active at time t_1 but not any more at time t_2, whereas $W_{t_1\cap t_2}$ corresponds to the traffic of the flows active at time t_1 and still active at time t_2. For clarity purposes, we introduce the following notations for the mean:

$$\lambda_W = \mathbb{E}\{W_{t_1\cap t_2}\} = \Lambda(C(t_1)\cap C(t_2)), \tag{5.6}$$

$$\mu_W = \mathbb{E}\{W_{t_1\backslash t_2}\} = \mathbb{E}\{W_{t_2\backslash t_1}\} = \Lambda(C(t_1)\backslash C(t_2)) = \Lambda(C(t_2)\backslash C(t_1)), \tag{5.7}$$

where simple calculations show that the last equality $\Lambda(C(t_1)\backslash C(t_2)) = \Lambda(C(t_2)\backslash C(t_1))$ holds due to the uniform distribution of the measure Λ in time. Finally, note that with these notations, we have:

$$\mathbb{E}\{W_{t_1}\} = \mathbb{E}\{W_{t_2}\} = \lambda + \mu. \tag{5.8}$$

The setting described here allows us to compute the autocovariance of the instantaneous bandwidth W_t, in order to evaluate the Hurst parameter, as we shall see now.

Computation of the instantaneous bandwidth correlation

We are interested in the computation of the autocovariance function of the instantaneous bandwidth, that is $\mathbb{E}\{W_{t_1}W_{t_2}\} - \mathbb{E}\{W_{t_1}\}\mathbb{E}\{W_{t_2}\}$.

Proposition 5.3.1. *If the planar point process $\{(T_i, D_i)\}$ is Poisson, then*

$$\mathbb{E}\{W_{t_1}W_{t_2}\} - \mathbb{E}\{W_{t_1}\}\mathbb{E}\{W_{t_2}\} = \mathbb{E}\{W^2_{t_1\cap t_2}\} - \mathbb{E}\{W_{t_1\cap t_2}\}^2 = \mathbb{V}ar\{W_{t_1\cap t_2}\} \tag{5.9}$$

Proof. Using the notations introduced above, we have:

$$\mathbb{E}\{W_{t_1}W_{t_2}\} = \mathbb{E}\{W_{t_1\backslash t_2}W_{t_2\backslash t_1}\} + \mathbb{E}\{W_{t_1\backslash t_2}W_{t_1\cap t_2}\} + \mathbb{E}\{W_{t_1\cap t_2}W_{t_2\backslash t_1}\} + \mathbb{E}\{W^2_{t_1\cap t_2}\}.$$

Due to the Poisson assumption, the random variables $W_{t_1\backslash t_2}, W_{t_2\backslash t_1}, W_{t_1\cap t_2}$ are mutually independent, so that

$$\begin{aligned}
\mathbb{E}\{W_{t_1}W_{t_2}\} &= \mathbb{E}\{W_{t_1\backslash t_2}\}\mathbb{E}\{W_{t_2\backslash t_1}\} + \mathbb{E}\{W_{t_1\backslash t_2}\}\mathbb{E}\{W_{t_1\cap t_2}\} + \mathbb{E}\{W_{t_1\cap t_2}\}\mathbb{E}\{W_{t_2\backslash t_1}\} + \mathbb{E}\{W^2_{t_1\cap t_2}\} \\
&= \mu_W^2 + 2\lambda_W\mu_W + \mathbb{E}\{W^2_{t_1\cap t_2}\} \\
&= (\lambda_W + \mu_W)^2 - \lambda_W^2 + \mathbb{E}\{W^2_{t_1\cap t_2}\}
\end{aligned}$$

directly giving equation (5.9) in view of the definition of λ_W (equation (5.6)) and of equation (5.8). □

Proposition 5.3.1 shows that the autocovariance depends only on the variance of the traffic due to the flows in the intersection of the cones $C(t_1)$ and $C(t_2)$. This had been noticed already in [Barral and Mandelbrot, 2002, Chainais et al., 2005], though the authors were focusing on small-scale properties. To complete the computation of the autocovariance function, we then need to compute $\mathbb{V}ar\{W_{t_1\cap t_2}\}$.

If the flow rates were constant and equal to 1, then $W_{t_1\cap t_2}$ would simply be the number of points in $C(t_1)\cap C(t_2)$. Since the point process is Poisson, the variance $\mathbb{V}ar\{W_{t_1\cap t_2}\}$ would then simply be λ_W and the autocovariance would be determined by the value of λ_W (recall that the variance of the count process associated with a Poisson process is equal to its mean). Before proceeding with the calculations in a more general case, we introduce additional useful notations. We denote by N the random variable corresponding to the number of points in $C(t_1)\cap C(t_2)$. We denote its mean by

$$\lambda_N = \mathbb{E}\{N\}. \tag{5.10}$$

Note that in the case where the rate is always 1, we simply have $\lambda_W = \lambda_N$. The next proposition gives the general form of the autocovariance function, without specifying the correlation between R_i and D_i yet.

Proposition 5.3.2.

$$\mathbb{E}\{W_{t_1}W_{t_2}\} - \mathbb{E}\{W_{t_1}\}\mathbb{E}\{W_{t_2}\} = \lambda_N\mathbb{E}\{R_i^2\}, \tag{5.11}$$

where it is implicitly understood that the expectation is computed in $C(t_1)\cap C(t_2)$.

Proof. To evaluate the value of $\mathbb{V}ar\{W_{t_1\cap t_2}\}$, we successively compute the values of $\mathbb{E}\{W_{t_1\cap t_2}\}$ and $\mathbb{E}\{W^2_{t_1\cap t_2}\}$.

$$\begin{aligned}
\mathbb{E}\{W_{t_1\cap t_2}\} &= \mathbb{E}\{\mathbb{E}\{W_{t_1\cap t_2}|N\}\} \\
&= \sum_{k=1}^{\infty}\mathbb{E}\{\sum_{i=1}^{k}R_i|N=k\}\mathbb{P}(N=k) \\
&= \sum_{k=1}^{\infty}\sum_{i=1}^{k}\mathbb{E}\{R_i|N=k\}\mathbb{P}(N=k) \\
&= \sum_{k=1}^{\infty}k\mathbb{E}\{R_i\}\mathbb{P}(N=k) \\
&= \lambda_N\mathbb{E}\{R_i\}.
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\{W^2_{t_1\cap t_2}\} &= \mathbb{E}\{\mathbb{E}\{W^2_{t_1\cap t_2}|N\}\} \\
&= \sum_{k=1}^{\infty}\mathbb{E}\{(\sum_{i=1}^{k}R_i)^2|N=k\}\mathbb{P}(N=k)
\end{aligned}$$

By independence of the sequence $(R_i)_i$

$$\mathbb{E}\{(\sum_{i=1}^{k}R_i)^2|N=k\} = \mathbb{E}\{(\sum_{i=1}^{k}R_i)^2\} = k\mathbb{E}\{R_i^2\} + (k^2-k)\mathbb{E}\{R_i\}^2,$$

so that

$$\begin{aligned}
\mathbb{E}\{W^2_{t_1\cap t_2}\} &= \sum_{k=1}^{\infty}(k\mathbb{E}\{R_i^2\}+(k^2-k)\mathbb{E}\{R_i\}^2)\mathbb{P}(N=k)\\
&= \sum_{k=1}^{\infty}k\mathbb{E}\{R_i^2\}\mathbb{P}(N=k)+\sum_{k=1}^{\infty}k^2\mathbb{E}\{R_i\}^2\mathbb{P}(N=k)-\sum_{k=1}^{\infty}k\mathbb{E}\{R_i\}^2\mathbb{P}(N=k)\\
&= \lambda_N\mathbb{E}\{R_i^2\}+(\lambda_N+\lambda_N^2)\mathbb{E}\{R_i\}^2-\lambda_N\mathbb{E}\{R_i\}^2\\
&= \lambda_N\mathbb{E}\{R_i^2\}+\lambda_N^2\mathbb{E}\{R_i\}^2
\end{aligned}$$

□

Until that point, we still have not used the precise form of the measure Λ (equation (5.2)), and the precise form of the correlation between R_i and D_i. The result of Proposition 5.3.2 depends only on the Poisson and independence of the sequence $(R_i)_i$ assumptions. It shows that the autocovariance function is the product of two terms: λ_N, the mean number of points in $C(t_1)\cap C(t_2)$, and $\mathbb{E}\{R_i^2\}$. The first term (λ_N) depends only on the measure Λ and is easily obtained via a simple integration (Proposition 5.3.3). To compute the second term ($\mathbb{E}\{R_i^2\}$), we need in addition to precise the correlation between R_i and D_i (Proposition 5.3.4).

Proposition 5.3.3. *If the measure Λ has the form of equation (5.2), then*

$$\lambda_N = C\frac{1}{\alpha_{\mathrm{ON}}(\alpha_{\mathrm{ON}}-1)}(t_2-t_1)^{-\alpha_{\mathrm{ON}}+1} \tag{5.12}$$

Proof.

$$\begin{aligned}
\lambda_N &= \int_{d=t_2-t_1}^{\infty}\int_{t=t_1-(d-(t_2-t_1))}^{t_1}\Lambda(\mathrm{d}u,\mathrm{d}v)\\
&= C\int_{d=t_2-t_1}^{\infty}(d-(t_2-t_1))\frac{1}{d^{\alpha_{\mathrm{ON}}+1}}\mathrm{d}d\\
&= C\int_{d=t_2-t_1}^{\infty}\frac{1}{d^{\alpha_{\mathrm{ON}}}}\mathrm{d}d-C(t_2-t_1)\int_{d=t_2-t_1}^{\infty}\frac{1}{d^{\alpha_{\mathrm{ON}}+1}}\mathrm{d}d\\
&= C\frac{1}{\alpha_{\mathrm{ON}}-1}(t_2-t_1)^{-\alpha_{\mathrm{ON}}+1}-C\frac{1}{\alpha_{\mathrm{ON}}}(t_2-t_1)^{-\alpha_{\mathrm{ON}}+1}\\
&= C\frac{1}{\alpha_{\mathrm{ON}}(\alpha_{\mathrm{ON}}-1)}(t_2-t_1)^{-\alpha_{\mathrm{ON}}+1}
\end{aligned}$$

□

Finally, we now specify the correlation between R_i and D_i to calculate the term $\mathbb{E}\{R_i^2\}$. Our goal here is to specify a correlation which indeed leads to the different tail indices of the flow-size and flow-duration distributions. We have already mentioned that taking R_i as a random variable independent of D_i would lead to the same tail indices, independently of the distribution of R_i, provided that its mean is finite (see for instance [D'Auria and Resnick, 2006, D'Auria and Resnick, 2008]). The different tail indices α_{SI} and α_{ON} can then only come from correlation between R_i and D_i. The simplest choice would be to deterministically take for each flow: $R_i = D_i^{\beta-1}$, where $\beta = \frac{\alpha_{\mathrm{ON}}}{\alpha_{\mathrm{SI}}}$. In this case, we would

have $S_i = D_i^\beta$ or equivalently $D_i = S_i^{1/\beta}$. This would effectively lead to the different tail indices α_{ON} and α_{SI} for the flow duration and size distributions. However, this assumption of a deterministic rate for a flow of a given duration is not realistic. Instead we choose a model where the conditional expectation and variance of the rate given the duration are specified. This model is given in the next proposition.

Proposition 5.3.4. *Assume that* $\mathbb{E}\{R_i|D_i\} = KD_i^{\beta-1}$, *where* K *is a constant and* $\mathbb{V}ar\{R_i|D_i\} = V$, *where* V *is a constant. We denote*

$$\alpha' = \alpha_{\mathrm{ON}} - 2(\beta - 1). \tag{5.13}$$

If $\alpha' > 1$, *then*

$$\mathbb{E}\{R_i^2\} = \frac{1}{\lambda_N} CK^2 \frac{1}{\alpha'(\alpha'-1)} (t_2 - t_1)^{-\alpha'+1} + V. \tag{5.14}$$

Proof.

$$\mathbb{E}\{R_i^2\} = \mathbb{E}\{\mathbb{E}\{R_i^2|D_i\}\} = K^2\mathbb{E}\{D_i^{2(\beta-1)}\} + \mathbb{E}\{\mathbb{V}ar\{R_i|D_i\}\}$$

The same kind of integration as for the proof of Proposition 5.12 give

$$\mathbb{E}\{D_i^{2(\beta-1)}\} = \frac{1}{\lambda_N} C \frac{1}{\alpha'(\alpha'-1)} (t_2 - t_1)^{-\alpha'+1},$$

while we clearly have $\mathbb{E}\{\mathbb{V}ar\{R_i|D_i\}\} = V$, completing the proof. □

We are now able to state the final proposition establishing the decrease of the autocovariance function with $t_2 - t_1$, and then the long-range dependence of the process $(W_t)_t$.

Proposition 5.3.5 (Autocorrelation function and long-range dependence of the process $(W_t)_t$)**.** *If* $\mathbb{E}\{R_i|D_i\} = KD_i^{\beta-1}$ *and* $\mathbb{V}ar\{R_i|D_i\} = V$, *where* K, V *are constants and* $\alpha' = \alpha_{\mathrm{ON}} - 2(\beta-1) > 1$, *then*

$$\mathbb{E}\{W_{t_1}W_{t_2}\} - \mathbb{E}\{W_{t_1}\}\mathbb{E}\{W_{t_2}\} = CK^2 \frac{1}{\alpha'(\alpha'-1)} (t_2-t_1)^{-\alpha'+1} + CV \frac{1}{\alpha_{\mathrm{ON}}(\alpha_{\mathrm{ON}}-1)} (t_2-t_1)^{-\alpha_{\mathrm{ON}}+1}. \tag{5.15}$$

The process $(W_t)_t$ *is then (asymptotically) long-range dependent with Hurst parameter* H *following equation (5.1) (*$H = \frac{3-\alpha_{\mathrm{H}}}{2}$*), where*

$$\alpha_{\mathrm{H}} = \min(\alpha', \alpha_{\mathrm{ON}}). \tag{5.16}$$

Proof. The first part is immediate from Propositions 5.3.2, 5.3.3 and 5.3.4. For the Hurst parameter, recall that (see Chapter 2) a process is long-range dependent of Hurst parameter H if its autocovariance function algebraically decreases like $(t_2 - t_1)^{2H-2}$. □

Proposition 5.3.5 is our main result for this section. It establishes the long-range dependence of the instantaneous bandwidth and gives the Hurst parameter. Before going back to the in2p3 trace and verifying the consistence of our model choice with the data, we make a few general remarks on the result of Proposition 5.3.5, and the choice of the model.

Let us first comment on specific values of $\beta = \frac{\alpha_{\mathrm{ON}}}{\alpha_{\mathrm{SI}}}$.

If $\beta = 1$ ($\alpha_{\mathrm{SI}} = \alpha_{\mathrm{ON}}$): This case corresponds to the classical case of Taqqu's model and other infinite source Poisson models, where the rate is not correlated to the duration and we have the same tail indices for the flow-size and flow-duration distributions. In this case, the result of Proposition 5.3.5 reduces to Taqqu's relation (5.1), where the tail index controlling the long-range dependence α_{H} is the unique tail index $\alpha_{\mathrm{SI}} = \alpha_{\mathrm{ON}}$ (bear in mind that there are no heavy-tailed OFF times here because we used a Poisson flow arrival).

If $\beta > 1$ ($\alpha_{\mathrm{SI}} < \alpha_{\mathrm{ON}}$): This is the case of the in2p3 trace, where the rate increases in average with the duration of the flows. In this case the tail index controlling the long-range dependence is $\alpha_{\mathrm{H}} = \alpha' < \alpha_{\mathrm{ON}}$. The long-term correlations are then stronger than in the case of constant rates. Note that, depending on the value of β, we can have either $\alpha_{\mathrm{H}} < \alpha_{\mathrm{SI}}$, $\alpha_{\mathrm{H}} > \alpha_{\mathrm{SI}}$ or $\alpha_{\mathrm{H}} = \alpha_{\mathrm{SI}}$. It means that the tail index controlling the long-range dependence does not necessarily lie between α_{SI} and α_{ON}, but can be smaller than both these tail indices.

If $\beta < 1$ ($\alpha_{\mathrm{ON}} < \alpha_{\mathrm{SI}}$): This case corresponds to a situation where the rate decreases in average with the duration of the flows. This could happen for instance with some scheduling policy giving priority to small flows. In this case the tail index controlling the long-range dependence is $\alpha_{\mathrm{H}} = \alpha_{\mathrm{ON}} < \alpha'$.

In the case where $V = 0$ (this corresponds to the deterministic case $R_i = D_i^{\beta-1}$), the tail index controlling the long-range dependence is always $\alpha_{\mathrm{H}} = \alpha'$, even if $\beta < 1$. Note that depending on the values of β, we can have in this case surprising situations similar to the situations discussed for $\beta > 1$ where the tail index controlling the long-range dependence α_{H} is greater than the two tail indices α_{SI} and α_{ON}. Note that the calculation of the autocovariance function of equation (5.15) could easily be performed with some other (non-constant) forms of the conditional variance $\mathbb{V}ar\{R_i|D_i\}$. For example, an algebraically increasing variance ($\mathbb{V}ar\{R_i|D_i\} = D_i^{\gamma}$) would lead to another index $\alpha'' = \alpha_{\mathrm{ON}} - \gamma$, possibly controlling the long-range dependence for some set of parameters.

The long-range dependence that we have discussed until now is asymptotic long-range dependence. However, depending on the values of the constants K and V, we can observe situations where within some intermediate range of scale, the term of the autocovariance function (5.15) with the larger exponent dominates. For example, in the case $\beta > 1$, the first term of the autocovariance asymptotically dominates, but if the value of V is very large and the value of K is not too large, the second term can dominate within some scale range, thus leading to *pseudo long-range dependence* controlled by the tail index α_{ON}. Note that the two constants K and V do not have the same units and cannot be compared directly to one another. Instead, the whole terms of equation (5.15) have to be compared.

Let us finally mention that to "specify enough" the correlation between the flow rates and durations, in order to compute the autocovariance function of equation (5.15), we had to specify only the first two conditional moments $\mathbb{E}\{R_i|D_i\}$ and $\mathbb{V}ar\{R_i|D_i\}$. This is not surprising since the autocovariance function is a second-order moment quantity, and we found only second-order self-similarity (*i.e.*, long-range dependence). To investigate finer statistical properties of the instantaneous bandwidth, we might need to specify the entire conditional probability $\mathbb{P}(R_i = r|D_i = d)$.

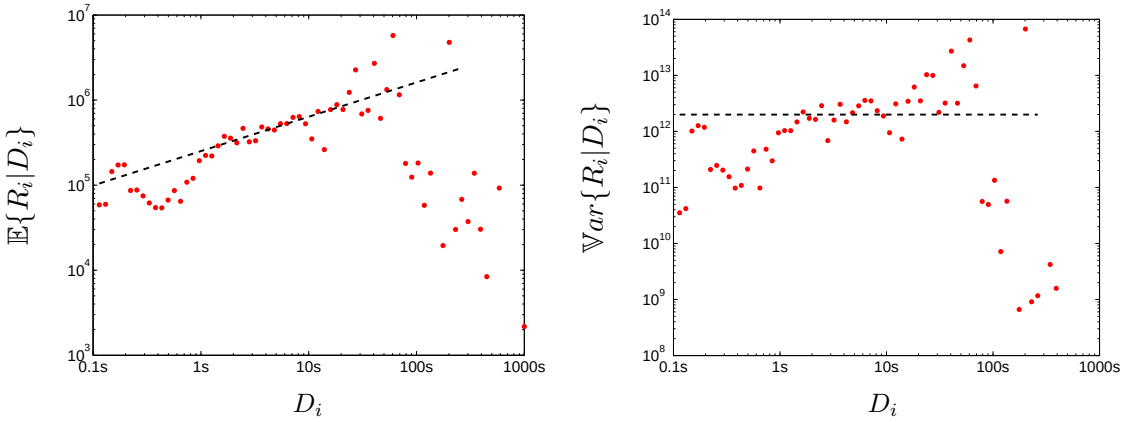

Figure 5.18: First two moments of the conditional probability of the flow rates given the flow durations. The red dots are empirical histograms estimated on the in2p3 trace. The black lines are not estimated from the empirical histograms. They have slope (left) $\beta - 1 = 0.4037$ where β is estimated from the tail indices α_{SI} and α_{ON} – (right) slope 0. They have been vertically adjusted to fit the data in the range $[1 - 100]$ s

5.3.3 Confrontation of the model with the real trace

We now come back to the in2p3 trace introduced in Section 5.3.1. Our goal is to confront the model we proposed in previous section to the real data, especially regarding the assumptions we made on the correlation between flow rates and durations in Proposition 5.3.5. From the in2p3 trace, we have obtained the estimations $\alpha_{\mathrm{SI}} = 0.8544$ and $\alpha_{\mathrm{ON}} = 1.1994$, which gives $\beta = 1.4037$, $\alpha' = 0.392$. Note that $\alpha' < 1$, so that the autocovariance function calculated in equation (5.15) theoretically diverges. As we shall see, the long-range dependence observed in Section 5.3.1 is in fact *pseudo long-range dependence*, *i.e.*, we are interested in the autocovariance function decrease in a finite-scale range, so that this divergence is not a problem (let us recall here that in Section 5.3.1 we have estimated the values $\widehat{H} = 0.901 \pm 0.045$ in the scale range: $[0.64 - 40.96]$ s).

Figure 5.18 shows the conditional moments $\mathbb{E}\{R_i|D_i\}$ and $\mathbb{V}ar\{R_i|D_i\}$, as functions of D_i, estimated from the in2p3 trace. We plotted lines corresponding to the model of Proposition 5.3.5: $\mathbb{E}\{R_i|D_i\} = KD_i^{\beta-1}$ and $\mathbb{V}ar\{R_i|D_i\} = V$, where K, V are constants, with the estimated value of β. These lines are vertically adjusted and show a good match in the range $[1 - 100]$ s, with the values $K = 10^{5.2}$, $V = 10^{12.4}$. This confirms the relevance of our model in this scale range, which roughly corresponds to the scale range in which we estimated the Hurst parameter in Section 5.3.1. With these values of K and V, we can compare the two terms of the autocovariance function of equation (5.15). We find that in the scale range of estimation of H ($[0.64 - 40.96]$ s), the first term of equation (5.15) is always five times smaller than the second term. We conclude that the long-range dependence observed on Figure 5.16 is pseudo long-range dependence controlled by the second term of equation (5.15), *i.e.*, by the tail index of the flow-duration distribution α_{ON}. The estimated value of the Hurst parameter shows a very good agreement with relation (5.1) with $\alpha_{\mathrm{H}} = \alpha_{\mathrm{ON}}$, confirming the accuracy of the model in Proposition 5.3.5 in predicting the Hurst parameter's value provided that the terms of equation (5.15) are appropriately compared in the scale range of interest.

5.3.4 Conclusion

In this section, we developed a model taking into account the correlation between flow rates and durations responsible for the different tail indices α_{SI} and α_{ON}. Based only on specifications of the first two order moments of the conditional probability of the rates given the duration, we showed that the instantaneous bandwidth exhibits long-range dependence with a Hurst parameter as in relation (5.1), where the controlling tail index α_{H} can be either α_{ON} or a combination of α_{ON} and α_{SI}, whichever is the smaller. We also showed that, in certain circumstances, pseudo long-range dependence with a different Hurst parameter can be observed in a finite scale range. Finally, we validated the ability of our model to predict the correct Hurst parameter on a real web traffic trace. However, the non-stationarity inherent to real Internet data at large time scales (around one hour) constrained us to restrain our study to 30 minutes of traces. Consequently, we observed pseudo long-range dependence, and longer stationary traces would be required to illustrate the ability of our model to predict asymptotic long-range dependence, which, as we saw, can exhibit a different Hurst parameter.

The model proposed in this section is the first model, to the best of our knowledge, that includes the correlation between flow rates and durations. This correlation is very important. It has been observed for a decade [Crovella and Bestavros, 1997], and is likely to become even more important with the appearance of FTTH and flow-aware approaches. In the case where this correlation vanishes, our results coincide with the usual results of Taqqu's theorem. Further theoretical developments would be needed to give a more rigorous mathematical statement of our results. Also, we assumed a Poisson flow arrival, which is a useful simplification to conduct the calculations, but might not always hold. Considering less restrictive flow-arrival processes would also be an interesting improvement of our results, be it only to clarify in which situations a correlated flow-arrival process can affect the aggregate traffic self-similarity.

Conclusion

In this chapter, we tackled the statistical characterization of aggregate network traffic, focusing on large-scale scaling properties, namely long-range dependence.

We first performed a thorough investigation of Taqqu's relation (equation (5.1)), in the loss-free situation, then in more realistic conditions implying losses. This study showed that, while the protocol TCP cannot be solely invoked to explain the long-range dependence effects, its mechanisms can affect the Hurst parameter in two ways: exponential backoff periods, splitting flows can annihilate the long-range dependence; and the RTT-scale correlation between the sources can induce a deviation as compared to Taqqu's relation. Beyond the proper use of robust statistical estimators, the very clear results that we obtained were enabled by the use of long-duration and controlled traces generated on our lar-scale experimental platform.

Moving beyond Taqqu's model, we then proposed a extension of this model taking into account unavoidable correlations between flow rates and durations. Our generalization is fully consistent with Taqqu's results (and other similar results), when this correlation falls down to zero. Our approach was based on a planar Poisson point process, a setting rarely used in long-range dependent models, which allowed us to refine the prediction of the Hurst parameter. We believe that this setting, and the correlation parameter we introduced, will

also conduct to interesting results on the small scales properties of the aggregate traffic.

Aggregate network traffic characterization.

- To what extend is Taqqu's relation valid in real network conditions? What are the scales involved? Does the protocol play a role? In which situation (loss-free, lossy link, congestion)?

In loss-free situations, Taqqu's relation is observed with a good agreement, independently of the protocol, provided that the long-range dependence is measured in a significant scale range lying far beyond the mean flow duration.

In the lossy-link case, the long-range dependence persists with TCP for small loss rates. For high loss rates, it breaks down because flows are split (by exponential backoff periods) into smaller flows whose size distribution appears exponential.

In the congestion case, the long-range dependence is present in sub-traffic corresponding to a part of the sources, but its Hurst parameter is affected by the correlation introduced by the interation of the sources at the buffer.

- What happens if flow-duration and flow-size distributions have different tail indices? Which of these tail indices, if any, governs the long-range dependence of the aggregate traffic?

A relation similar to Taqqu's relation still holds. The governing tail index can be the α_{ON} or a combination of α_{ON} and α_{SI}, whichever is the smaller. In some finite scale range, the larger of these indices can dominate in some circumstances.

Chapter 6

Estimation of the flow-size distribution's tail index from sampled data

The results of this chapter have been published in [3, 5].

Source-traffic characterization at flow scale.

- How to estimate the flow-size distribution's tail index from packet-sampled data?

Introduction

In the previous chapter, we assumed that the flow-size distribution had a given tail index, and gave answers to the question of characterizing the aggregate traffic. We now turn to the opposite question of characterizing the source traffic at flow scale, and focus on the estimation of the flow-size distribution's tail index. As we have seen in the previous chapter that this tail index is likely to control the long-range dependence in many situations, its estimation is indeed of primary importance. Moreover, knowledge of this tail index, as an important part of the complete system description, can be very usefull in problems such as characterization of the network's resource usage, dimensioning, and scheduling.

Most proposed methods for estimating the tail index (see *e.g.* [Seal, 1952, Hill, 1975, Crovella and Taqqu, 1999, Nolan, 2001, Gonçalves and Riedi, 2005, Clauset et al., 2009]) rely on observing the full traffic, implying the capture of every packets. However, with very-high-speed networks, these methods become very demanding in terms of memory resources and CPU consumption. It is thus necessary to sample the packet stream, retaining (deterministically or randomly) only a subpart of the aggregated traffic going through the link. This raises the question of how to estimate the flow-size distribution's tail index from sampled data, that we address in this chapter.

The more general problem of inferring the original flow-size distribution from sampled data has been extensively studied in the literature [Duffield et al., 2003, Hohn and Veitch, 2006, Liu et al., 2006, Ribeiro et al., 2006a, Yang and Michailidis, 2007, Chabchoub et al., 2009]. However, and despite its relatively simple formulation, no conclusive estimation procedure or closed-form estimate expression has been derived so far. All approaches undertake simplifying assumptions that lead to approximate solutions only.

In this chapter, we derive the exact solution of the maximum-likelihood problem in the case where the original flow-size distribution is a zeta distribution, and provide an analytic expression for the heavy-tail exponent's estimate from a sub-sampled packet series. We first expose in Section 6.1 the mathematical formulation of the problem and set the definitions and notation. Section 6.2 lists some of the previously proposed solutions to the problem of tail-index estimation from sampled observations. In Section 6.3, we state and explicitly solve the maximum-likelihood formulation of the problem. The proposed MLE solution is given an enlightening interpretation in terms of a geometric solution that we proposed in a former work. We complete this section with the Cramér-Rao bound derivation. Section 6.4 shows the results of numerical simulations to assess and rate MLE performance with respect to that of previous approaches; and then compares the performance of the MLE to the performance of the other estimators on the basis of a real Internet traffic trace.

Before developing our solution for the specific problem of estimating the flow-size distribution's tail index, let us mention that the impact of sampling has recently been studied in many other contexts (this list of papers is not exhaustive and the interested reader can see the references within the papers cited here): in [Mori et al., 2004] and [Estan and Varghese, 2002], the authors tackle the problem of identifying elephant flows from sampled data. In [Barakat et al., 2005], the authors consider the problem of ranking the largest flows on a link under packet sampling. In [Brauckhoff et al., 2006] and [Kawahara et al., 2007], the authors study the impact of sampling on anomaly-detection methods. Furthermore, in [Hohn and Veitch, 2006], the authors point out that *flow sampling* (*i.e.*, the sampling decision is made on the flows, and every packet of a picked flow is collected) is a way to bypass the major difficulties inherent to packet sampling in the flow statistics' estimation. As a way to combine the good statistical properties of *flow sampling* with the low computational cost of *packet sampling*, the authors in [Tune and Veitch, 2008] propose *Dual Sampling*. However, as it remains the most simple sampling procedure, and as it is implemented in many routers [13], we will only consider packet sampling in this work.

6.1 Problem definition and notation

In this paper, we analyze traffic time series consisting of a succession of TCP-IP packets, observed over a time duration T. Recall that, while the strict notion of flow still fuels an active debate, we adopt the consensual definition that flows are reconstituted by regrouping packets sharing the same protocol, the same source and destination IP addresses, and the same source and destination port, and that the flow size is then defined as the number of packets therein. It is a discrete random variable that we denote S. The original flow-size distribution reads:

$$P_S(S = i) = \phi_i,\ i \in \mathbb{N}. \tag{6.1}$$

It can be empirically estimated as the normalized frequency of flows with size i that are observed over the period T.

For discrete random variables, the zeta distribution is the paradigm of heavy-tailed distributions (see equation (2.7) for the definition of heavy-tailed distributions):

$$P(S = i; \alpha) = \frac{i^{-(\alpha+1)}}{\zeta(\alpha + 1)}; \tag{6.2}$$

where

$$\zeta(\alpha+1) = \sum_{k=1}^{\infty} k^{-(\alpha+1)} \tag{6.3}$$

is the Riemann zeta function. The zeta distribution is the discrete counterpart of the Pareto distribution. In our study, we implicitly consider original flows whose sizes are i.i.d. random variables drawn from a heavy-tailed distribution of the form (6.2). Our goal is then to estimate the tail exponent α of this underlying distribution.

Estimating the α index of heavy-tailed distributions from a set of independent realizations of the random variable S is a classical problem, broadly treated in the literature for both continuous and discrete variables. For instance, Hill [Hill, 1975] and Nolan [Nolan, 2001], proposed maximum-likelihood estimators (MLE) for Pareto and alpha-stable laws respectively, whereas Seal in [Seal, 1952], derived the discrete MLE counterpart for a zeta distribution, recently revisited in [Clauset et al., 2009]. The wavelet-based estimator of [Gonçalves and Riedi, 2005] presented in Section 2.4.2 and used in previous chapter also works in this situation for a general heavy-tailed distribution.

When variable S is not directly observable, but only a thinned version of it, things become more complicated, and the proposed solutions not so conclusive. That is precisely the case when memory and CPU consumption issues compel to sample the packets streams. Thus, the observed flow size, defined as the number of sampled packets within a flow, is a random variable Y on its own, different from the original flow size S, and following the new distribution:

$$P_Y(Y = j) = \eta_j. \tag{6.4}$$

The simplest way to practically perform packet sampling consists in sequentially picking one packet every $K \in \mathbb{N}$, where $p = \frac{1}{K}$ is the sampling rate. However, for theoretical development purposes, it is much simpler to consider random sampling, which consists in randomly picking every packet with a probability p. The two methods have been proven equivalent when a sufficiently large number of intertwined flows is assumed [Chabchoub et al., 2007]. Thereafter, a probabilistic sampling with sampling rate p is always assumed. Then, the conditional probability that a sampled flow of size j comes from an original flow of size $i \geq j$ is simply governed by a binomial law:

$$P_{Y|S}(Y = j|S = i) = B_p(i, j) = \binom{i}{j} p^j (1-p)^{i-j}, \tag{6.5}$$

From this, we get the sampled flow size distribution expressed in terms of the original flow size distribution:

$$P_Y(Y = j) = \eta_j = \sum_{i=j}^{\infty} B_p(i, j)\phi_i. \tag{6.6}$$

Let us notice that equation (6.6) is properly normalized provided that all sampled flow sizes, including $j = 0$, can be observed. As this is generally unrealistic in practice[1], we are

[1]With the TCP protocol, the number of missed flows can be estimated as the difference between the total number of expected flows, given by the counting of observed SYN packets divided by p, and the number of actually observed flows.

led to introduce the minimal observable size of a sampled flow: $j_{\min}$ (so that $P_Y(Y=j)=0$ if $j<j_{\min}$), and the correct re-normalization of the sampled flow size distribution is:

$$P_Y(Y=j)=\eta_j = \frac{\sum_{i=j}^{\infty} B_p(i,j)P_S(S=i)}{\sum_{j'=j_{\min}}^{\infty}\sum_{l=j'}^{\infty} B_p(l,j')P_S(S=l)} = \frac{\sum_{i=j}^{\infty} B_p(i,j)\phi_i}{\sum_{j'=j_{\min}}^{\infty}\sum_{l=j'}^{\infty} B_p(l,j')\phi_l}, \tag{6.7}$$

for all $j \geq j_{\min}$. The normalization factor naturally reduces to 1 when $j_{\min}=0$.

The problem we address in this paper is the following: how can we estimate the tail index α of the original flow-size distribution P_S from a finite set of sampled flow sizes observations $\{y_k\}_{k=1,\ldots,m}$?

6.2 Related work

In this Section, we provide a comprehensive survey of the major methods existing to estimate the flow-size distribution's tail index from sampled data, clarifying their different nature and their common features.

One can tackle this estimation issue from two distinct perspectives: (a) by applying a *standard* tail-index estimator to the flow-size distribution which has been priorly estimated from the sampled observations (2-steps methods); or (b) by directly inferring the tail index from the sampled observations (without estimating the underlying distribution). These approaches are described in Sections 6.2.1 and 6.2.2, respectively.

6.2.1 Two-steps methods

In the presence of fully observed data, tail-index estimation has been largely investigated and the 2-steps procedures share the same estimation choice given below (Section 6.2.1). Regarding the inference of the unknown flow-size distribution, several possibilities are listed in Section 6.2.1.

Tail-exponent estimation

Let us assume that we were able to properly estimate the original flow-size distribution beyond a minimal size $i_{\min}$ from the observed sampled flows. If this latter obeys a power-law model of the form (2.7), estimating the tail index α is not plagued by the sampling obstacle anymore, and the *classical* estimators mentioned above can directly apply. In this study, we choose to use a Hill-type method [Hill, 1975], whose good performance is notoriously acknowledged by the statistical community. The Hill estimator from the inferred original flow-size distribution $\{\widehat{\phi}_i\}_{i=i_{\min},\ldots,\infty}$ reads:

$$\widehat{\alpha}_{\text{Hill}} = \left(\sum_{i=i_{\min}}^{\infty} \widehat{\phi}_i \ln \frac{i}{i_{\min}} \right)^{-1}. \tag{6.8}$$

A slightly modified Hill-estimation procedure has been proposed in [Clauset et al., 2009] as a more appropriate procedure than the *classical* Hill estimator for discrete random

variables:

$$\widehat{\alpha}_{\text{Discrete}} = \left(\sum_{i=i_{\min}}^{\infty} \widehat{\phi}_i \ln \frac{i}{i_{\min} - \frac{1}{2}} \right)^{-1}. \tag{6.9}$$

Note that these two procedures lead to the same estimation if $i_{\min}$ is sufficiently large.

Next Section details some proposed methods to estimate the original flow-size distribution from a sampled packet sequence.

Flow-size distribution inference

Inverse approximation using an *a priori*.

The basic idea behind inverse approximation is to recover the original distribution via the sum and product rules:

$$P_S(S = i) = \sum_{j=0}^{\infty} P_{S|Y}(S = i|Y = j) P_Y(Y = j), \tag{6.10}$$

where the conditional probability $P_{S|Y}$ needs first to be estimated. To this end, we use the Bayes formula where we initialize the unknown distribution P_S to an *a priori* distribution P_S^{ap}, and get:

$$\begin{aligned} P_{S|Y}(S = i|Y = j) &= \frac{P_{Y|S}(Y = j|S = i) P_S^{ap}(S = i)}{P_Y(Y = j)} \\ &= \frac{B_p(i, j) P_S^{ap}(S = i)}{\sum_{l=j}^{\infty} B_p(l, j) P_S^{ap}(S = l)}. \end{aligned} \tag{6.11}$$

In this approach, the estimation accuracy of P_S essentially depends on a relevant choice of P_S^{ap}.

Uniform *a priori* with rectangular approximation of the conditional probability – scaling method. The simplest *a priori* distribution that we can plug in equation (6.11) corresponds to a uniform *a priori*: $P_S^{ap}(S = i) = C,\ \forall i$, which yields the conditional probability:

$$P_{S|Y}(S = i|Y = j) = \frac{B_p(i, j)}{\sum_{l=j}^{\infty} B_p(l, j)} = p \cdot B_p(i, j). \tag{6.12}$$

We can further simplify this expression, approximating the binomial function by a simplistic rectangular window:

$$\begin{aligned} P_{S|Y}(S = i|Y = j) &= p, \text{ for } i = \frac{j}{p} + 1, \ldots, \frac{j+1}{p} \\ &= 0, \text{ otherwise}, \end{aligned} \tag{6.13}$$

which finally leads to the original distribution estimate:

$$\widehat{\phi}_i = p\, \eta_{\lfloor ip \rfloor},\ \forall i. \tag{6.14}$$

This expression was originally proposed in [Duffield et al., 2003] and called *scaling estimator*, as it simply corresponds to the sampled flow-size distribution, re-scaled by a factor

p. Figure 6.1(a) shows an example of flow-size distribution inference with the scaling method. It illustrates the principal characteristic of this method: the inferred distribution is piecewise constant on intervals of length $1/p$ intervals (on Figure 6.1(a), the step width decreases with i because of the logarithmic x scale).

Zeta *a priori* with geometric mean approximation. As we are interested in heavy-tailed distributions, a more natural choice for the *a priori* distribution P_S^{ap} is the zeta distribution of equation (6.2) with pre-fixed tail index α^{ap}. Under this assumption, the conditional probability (6.11) becomes:

$$P_{S|Y}(S=i|Y=j) = \frac{B_p(i,j)i^{-(\alpha^{ap}+1)}}{\sum_{l=j}^{\infty} B_p(l,j)l^{-(\alpha^{ap}+1)}}. \tag{6.15}$$

Then, instead of distributing the binomial mass $B_p(i,j)$ uniformly on the interval $[jp^{-1}+1,(j+1)p^{-1}]$, we concentrate all the conditional probability mass corresponding to a given j, on a unique point denoted $\bar{\imath}_{(\alpha^{ap})}(j)$:

$$\begin{aligned} P_{X|Y}(X=i|Y=j) &= 1, \text{ for } i=\bar{\imath}_{(\alpha^{ap})}(j) \\ &= 0, \text{ otherwise.} \end{aligned} \tag{6.16}$$

We heuristically proposed in [5] to choose $\bar{\imath}_{(\alpha^{ap})}(j)$ as the geometric mean of the sequence $i=j,\ldots,\infty$, weighted by the conditional probability (6.15):

$$\bar{\imath}_{(\alpha^{ap})}(j) = \exp\left(\frac{\sum_{i=j}^{\infty}\ln(i)B_p(i,j)i^{-(\alpha^{ap}+1)}}{\sum_{l=j}^{\infty}B_p(l,j)l^{-(\alpha^{ap}+1)}}\right). \tag{6.17}$$

A possible justification for this approximation lies in the fact that geometric means are *naturally adapted* to hyperbolic functions such as the power-law decay of heavy-tailed distributions. In Section 6.3, we will demonstrate that the expression $\bar{\imath}$ consistently arises when deriving the exact solution of the maximum-likelihood estimator. Figure 6.1(b), displaying an example of inference with the zeta *a priori* method shows that the inferred distribution with this method is much more accurate than with the scaling method, and already gives an empirical justification for this approximation based on the definition of $\bar{\imath}$.

In contrast with the scaling method of [Duffield et al., 2003], the proposed zeta *a priori* depends on the tail exponent α^{ap} that is precisely to be estimated. We then suggest an iterative procedure which uses the tail index estimated at step $k-1$ to set the *a priori* α^{ap} of step k.

Maximum-likelihood estimation of the original distribution.

In [Duffield et al., 2003] also, Duffield et al. tackle the direct estimation of the ϕ_i's in equation (6.1), solving the maximum-likelihood formulation of the problem. However, as this approach is highly sensible to the variance of the observations (*i.e.*, to the sampled flow-size frequencies), they show that maximizing the likelihood function yields negative frequencies. Therefore, imposing a positive constraint on the ϕ_i's, they resorted to an iterative Expectation-Maximization (EM) algorithm [Dempster et al., 1977, McLachlan and Krishnan, 1997] whose output converges towards the ML solution. Notwithstanding its persuasive interest, the difficulty to define a relevant criterion to stop the EM iterations,

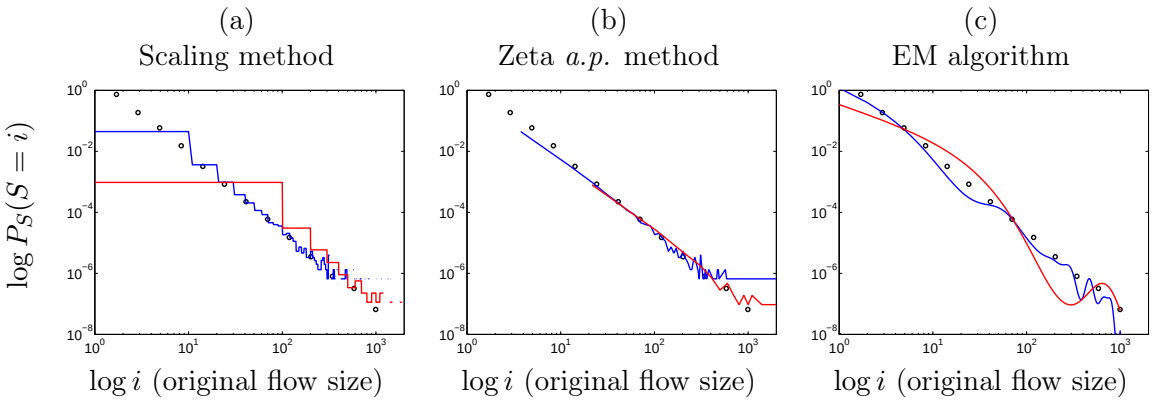

Figure 6.1: Original flow-size distribution inference with: (a) the scaling method – (b) the zeta *a priori* method – (c) the EM algorithm. The curves represent the inferred distributions with sampling rates: (blue) $p = 1/10$ – (red) $p = 1/100$. The black dots represent the original flow-size distribution (Zeta, $\alpha = 1.5$).

induces oscillating phenomena that drastically distort the tail decay of the inferred distribution, and lead to highly biased tail-index estimates (see Figure 6.1(c)). For this reason, we discard this method from our numerical evaluation in Section 6.4.

This approach from Duffield et al. has recently been improved in two different ways: in [Ribeiro et al., 2006a], the authors propose a method taking advantage of the protocol informations contained in a packet (mainly the TCP sequence numbers) to improve the accuracy of the estimation. In [Yang and Michailidis, 2007], the authors use a mixture model for the original flow-size distribution, separating small and large flows, and jointly estimate the mixture coefficient and the distribution. However, as we consider here a rigorous zeta distribution for the flow-size distribution, and we assume no knowledge about the protocol information, these methods fail to apply in our context.

Expansion of the probability-generating function.

In [Hohn and Veitch, 2006], the authors propose an original spectral approach to recover the probability densities of the original flow sizes from the thinned packets distributions. The central idea relies on the one-to-one correspondence between the density function and the probability generating function of a random variable. Then, due to the analyticity properties of the latter, theoretical results from complex analysis can directly apply to infer the original flow-size distribution from the probability generating function of the sampled packets' time series. Two distinct methods are devised: the first one constructs a power series expansion of the thinned probability generating function about the origin, and finds its non-trivial *analytic continuation* to the entire analytic domain of the original flow-sizes random variable. The second one is based on the *Cauchy integral formula*, whose evaluation on any closed contour including the origin can lead to the desired ϕ_i's.

According to the author's own evaluation, both methods perform fairly well only for values of the sampling rate $p > 0.5$. As we are interested in a sparser packet thinning (typically $p < 0.1$), we will not proceed with this approach. This limitation notwithstanding, its elegant theoretical formulation deserves to be stressed here.

6.2.2 Direct tail-index estimation

Although more straightforward, methods which do not imply prior estimation of the underlying heavy-tailed flow-size distribution (or characteristic function), but directly deduce the tail index α from the sampled packet series have received much less attention. To our knowledge, the only existing approach of this kind was proposed in [Chabchoub et al., 2008, Chabchoub et al., 2009] (see also [Chabchoub, 2009]).

The method relies on stochastic counting. Under the same sampling conditions, let W_k be a random variable defined as the number of sampled flows observed k times during a given observation period Δ. Using a Poisson approximation, and assuming *(i)* that the total number of packets is much larger than p^{-1}, and *(ii)* that the number of flows is large enough, the authors in [Chabchoub et al., 2008, Chabchoub et al., 2009] analytically prove that a relation between $\mathbb{E}\{W_k\}$ and $\mathbb{E}\{W_{k+1}\}$ holds, leading to the following estimate of α:

$$\widehat{\alpha}_k = (k+1)\left(1 - \frac{\mathbb{E}\{W_{k+1}\}}{\mathbb{E}\{W_k\}}\right) - 1, \text{ for } k \geq k_0. \tag{6.18}$$

In this expression, k_0 is a threshold defined to ensure that the counting process only comprises flow sizes lying in the power-law decay of the distribution, and beyond which $\widehat{\alpha}_k$ converges to the expected value of the tail index.

In practice, the duration Δ is divided into M non-overlapping shorter time intervals of size Δ/M. For each segment $m = 1, \ldots, M$, the counts $W_k^{(m)}$ simulate independent realizations of the random variable W_k. The empirical mean $M^{-1}\sum_{m=1}^{M} W_k^{(m)}$ then substitutes the ensemble average $\mathbb{E}\{W_k\}$ in (6.18).

6.3 Maximum-Likelihood Estimation of the tail index

We now elaborate on the direct estimation of the tail index α, from a statistical angle.

6.3.1 Formulation

In what follows, we assume that the original flow-size distribution follows a heavy-tailed zeta distribution of the form (6.2). Under this condition, the sampled flow-size distribution (6.6) becomes:

$$P_Y(Y = j|\alpha) = \frac{1}{\zeta(\alpha+1)} \sum_{i=j}^{\infty} B_p(i,j) i^{-(\alpha+1)}, \tag{6.19}$$

where all sampled flow sizes, including $j = 0$, are implicitly observed ($j_{\min} = 0$). Then, bypassing the unstable estimation of the underlying original flow-size distribution P_S, we can directly express the log-likelihood function as:

$$\mathcal{L}(\alpha) = -n \ln \zeta(\alpha+1) + n \sum_{j=0}^{\infty} \eta_j \ln \left(\sum_{i=j}^{\infty} B_p(i,j) i^{-(\alpha+1)} \right), \tag{6.20}$$

where n is the number of observed sampled flow sizes. In practice, when not all the sampled flow sizes are observed (*i.e.* $j_{\min} > 0$), it is the properly normalized form (6.7) that needs to be adopted and accordingly, equations (6.19) and (6.20) take on the form given in Appendix A, Section A.1.

Formally, the maximum-likelihood estimate of the tail index α is a solution to the following maximization problem:

$$\widehat{\alpha_{\mathrm{ML}}} = \underset{\alpha}{\operatorname{argmax}}\, \mathcal{L}(\alpha), \tag{6.21}$$

and is asymptotically unbiased.

6.3.2 Resolution and interpretation

Differentiating the log-likelihood function (6.20) and equaling the result to zero readily brings out the quantity $\overline{\imath}_{(\alpha)}(j)$ that we intuitively introduced in [5] (see Section 6.2.1):

$$\frac{\zeta'(\alpha+1)}{\zeta(\alpha+1)} = -\sum_{j=0}^{\infty} \eta_j \ln \overline{\imath}_{(\alpha)}(j). \tag{6.22}$$

A closed-form solution to this equation is not available and we adopted a fixed-point resolution technique. We numerically checked that such a fixed-point iteration converges toward the maximum-likelihood estimate $\widehat{\alpha_{\mathrm{ML}}}$ within a reasonable number of iterations (see Section 6.4.1). In addition, following the lines of [Duffield et al., 2003], we found that solving this ML problem via an Expectation-Maximization algorithm leads to the exact same iterated procedure, whose solution then coincides with the MLE (see Appendix A, Section A.3, were the derivation is presented).

This striking accordance draws an indirect justification for the geometric transformation of relation (6.17). Indeed, let us suppose that, regardless of the choice α^{ap}, the transformed variable $\overline{\imath}_{(\alpha^{ap})}(Y)$ does follow a heavy-tailed Pareto distribution with tail exponent α. This assertion can be justified as follows. Firstly, given that the random variable S follows a Pareto distribution with tail exponent α, the random variable Y is asymptotically heavy-tailed with the same exponent α. This is proven in two different ways in [Hohn and Veitch, 2006] and [Chabchoub et al., 2007]. In [Hohn and Veitch, 2006], the proof is based on the study of the generating functions of S and Y with a Tauberian Theorem (see [Bingham et al., 1987, p. 333]). In [Chabchoub et al., 2007], the proof is based on a Berry-Esséen Theorem (see [Feller, 1971, p. 542]). Then, given that the random variable Y is asymptotically heavy-tailed with tail exponent α, equation (A.11) (see Section A.2 of Appendix A) clearly shows that the transformed variable $\overline{\imath}_{(\alpha^{ap})}(Y)$ is also asymptotically heavy-tailed with the same exponent α. Experimentally, we checked that this heavy-tailedness of $\overline{\imath}_{(\alpha^{ap})}(Y)$ indeed holds when considering only the observations y_k larger than some threshold $j_{\min}$ rarely exceeding 3 or so. Under this hypothesis, maximization of the corresponding maximum-likelihood principle leads to the equation

$$\frac{\chi'(\alpha+1, j_{\min})}{\chi(\alpha+1, j_{\min})} = -\sum_{j=j_{\min}}^{\infty} \eta_j \ln \overline{\imath}_{(\alpha^{ap})}(j), \tag{6.23}$$

where

$$\chi(\alpha+1, j_{\min}) = \sum_{j=j_{\min}}^{\infty} \left(\overline{\imath}_{(\alpha^{ap})}(j)\right)^{-(\alpha+1)}. \tag{6.24}$$

In appendix A.2, we show that for large values of $j_{\min}$:

$$\chi(\alpha+1, j_{\min}) \simeq p \cdot \zeta(\alpha+1, \overline{\imath}_{(\alpha)}(j_{\min})),$$

where

$$\zeta(\alpha+1, i_{\min}) = \sum_{i=0}^{\infty} (i+i_{\min})^{-(\alpha+1)},$$

which clearly implies:

$$\frac{\chi'(\alpha+1, j_{\min})}{\chi(\alpha+1, j_{\min})} = \frac{\zeta'(\alpha+1, \overline{\imath}_{(\alpha)}(j_{\min}))}{\zeta(\alpha+1, \overline{\imath}_{(\alpha)}(j_{\min}))}.$$

This shows that applying directly a maximum-likelihood principle to the transformed random variable $\overline{\imath}_{(\alpha^{ap})}(Y)$ for the tail-index estimation with a heavy-tailed Pareto hypothesis (*i.e.* using equation (6.23), and iterating), leads to the exact same procedure as applying a fixed-point method to solve the initial maximum-likelihood problem (equation (6.22)).

In appendix A.2, we also show that for large values of $j_{\min}$:

$$\chi(\alpha+1, j_{\min}) \simeq p \cdot \frac{(\overline{\imath}_{(\alpha)}(j_{\min}))^{-\alpha}}{\alpha}. \tag{6.25}$$

Plugging equation (6.25) into equation (6.23) yields, after differentiation:

$$\widehat{\alpha} = \left(\sum_{j=j_{\min}}^{\infty} \eta_j \ln \frac{\overline{\imath}_{(\alpha^{ap})}(j)}{\overline{\imath}_{(\alpha^{ap})}(j_{\min})} \right)^{-1}, \tag{6.26}$$

which is a classical Hill estimation applied to the random variable $\overline{\imath}_{(\alpha^{ap})}(Y)$ (see equation (6.8)). Note that as previously mentioned, a classical Hill estimation (equation (6.8), [Hill, 1975]) and a modified estimation (equation (6.9), [Clauset et al., 2009]) lead to the same result because in our case the practical values of $\overline{\imath}_{(\alpha^{ap})}(j_{\min})$ are sufficiently large.

6.3.3 Properties of the MLE

One additional feature of our approach is that being statistically well-based, theoretical properties of the proposed estimator are accessible and allow us to evaluate its performance in terms of bias and variance.

As previously mentioned, the ML estimator is asymptotically unbiased. In addition, we can derive the theoretical Cramér-Rao bound fixing, for a given sample size, the minimum variance that an unbiased estimator can achieve, that is:

$$\mathbb{V}ar(\widehat{\alpha}) \geq \frac{1}{\mathcal{I}(\widehat{\alpha})}, \tag{6.27}$$

with Fisher information:

$$\mathcal{I}(\widehat{\alpha}) = -\mathbb{E}\left\{ \frac{\partial^2}{\partial \alpha^2} \mathcal{L}(\alpha) \right\} \Big|_{\alpha=\widehat{\alpha}}.$$

If the MLE is only asymptotically unbiased, the Cramér-Rao bound is the minimum variance only asymptotically. An unbiased estimator is said efficient if its variance attains the Cramér-Rao bound, and if such estimator exists, then the MLE is necessarily efficient.

In our case, straightforward differentiation of equation (6.20) gives

$$\mathcal{I}(\alpha) = n\left[\frac{\zeta''(\alpha+1)\zeta(\alpha+1)-\zeta'^2(\alpha+1)}{\zeta^2(\alpha+1)} + \mathbb{E}_Y\left\{\left(\frac{\sum_{i=1}^{\infty}\ln i B_p(i,Y)i^{-(\alpha+1)}}{\sum_{l=1}^{\infty}B_p(l,Y)l^{-(\alpha+1)}}\right)^2\right\}\right.$$
$$\left. -\mathbb{E}_Y\left\{\frac{\sum_{i=1}^{\infty}(\ln i)^2 B_p(i,Y)i^{-(\alpha+1)}}{\sum_{l=1}^{\infty}B_p(l,Y)l^{-(\alpha+1)}}\right\}\right], \tag{6.28}$$

where, again $j_{\min} = 0$ is implicit. A different value of $j_{\min}$ essentially modifies the first term of this sum, where the Riemann zeta functions undergo the same change as in equation (6.7). The equation corresponding to $j_{\min} > 0$ is given in Appendix A, Section A.1 (eq. (A.3)). In equation (A.3), both n and its multiplicative factor depend on $j_{\min}$. However, as illustrated by the plots of Figure 6.2, the decrease of n with $j_{\min}$ always dominates the non-monotonous variations of the other term. Therefore, the Cramér-Rao bound, as the Fisher information, is essentially controlled by n, the number of observed sampled flows.

Note that the minimal index in the summations have been set to 1 in these equations instead of the realization y_k of the random variable Y. Since $B_p(i,j)$ is equal to zero for $i < j$, the sum remains unchanged.

Notably, we find that the second term of the right-hand size of the sum simply reduces to $\mathbb{E}_Y\{\ln^2 \bar{\imath}_{(\alpha)}(Y)\}$, where $\bar{\imath}_{(\alpha)}$ was heuristically defined in (6.17).

6.4 Results

6.4.1 Performance evaluation using numerical simulations

Simulation scheme

We numerically evaluate the performances of the maximum-likelihood estimator $\widehat{\alpha_{\mathrm{ML}}}$ derived in previous section, and draw up a comparative study with the other estimators itemized in Section 6.2. Our study relies on synthetic traffic generated under matlab, as exposed in Section 4.4. This allows us to flexibly adjust the different influencing parameters such as the tail exponent of the prescribed flow-size distributions.

The traffic simulated reproduces the aggregated traffic generated by $N_{\mathrm{sources}} = 100$ homogeneous ON/OFF sources.

The flow-size distribution is prescribed to a zeta distribution (6.2), with tail index α. In addition, we fixed the following traffic characteristics:

- OFF periods are exponentially distributed. OFF and ON durations have the same mean.
- Each source rate is set to 10 Mbps, resulting in a mean aggregated traffic of 500 Mbps.
- Packet size is constant and fixed to 1500 Bytes.
- Experimental data corresponds to a stationary packets series generated over a $T =$ 300 s period.

The tail index α takes on five possible values: 1.1, 1.3, 1.5, 1.7 and 1.9. For each of those, we generated 50 independent time series, that we randomly thinned afterwards, imposing three different values for the sampling rate p: 1/10, 1/100 and 1/1000. For each combination of α and p, bias and variance of all studied estimators are empirically evaluated from the 50 independent realizations.

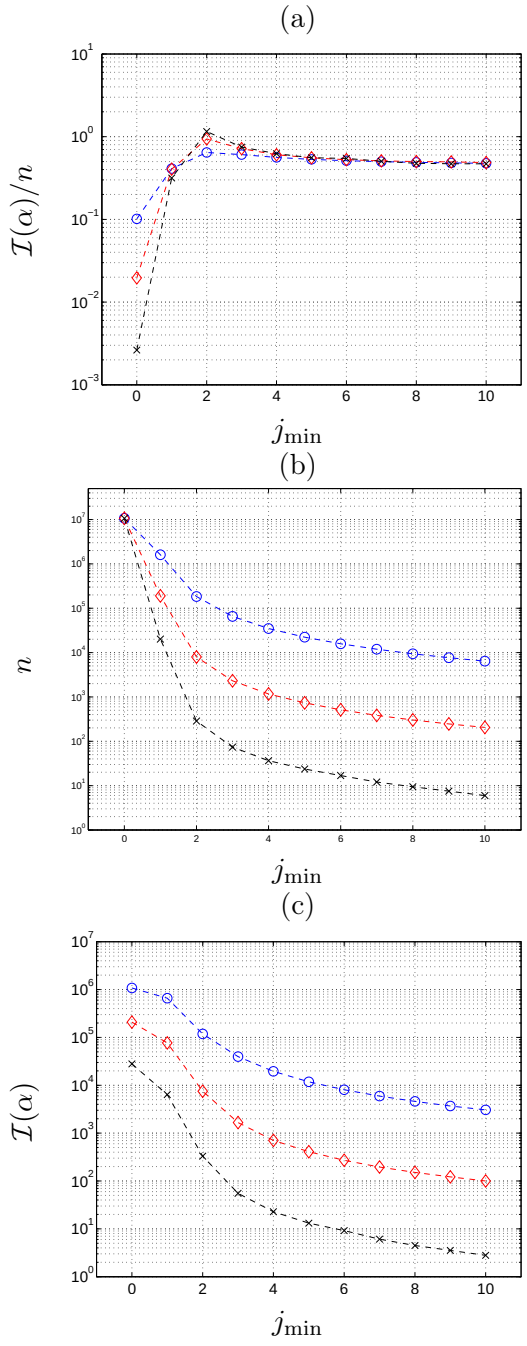

Figure 6.2: Variations of the different parts of the Fisher information with $j_{\min}$ (see eq. (A.3)): (a) the pre-factor of the Fisher information $\mathcal{I}(\alpha)/n$ – (b) the number of sampled flows n – (c) the total Fisher information $\mathcal{I}(\alpha)$. The three curves on each graph correspond to three different values of p: ($\bigcirc$, blue) $p = \frac{1}{10}$ – ($\Diamond$, red) $p = \frac{1}{100}$ – ($\times$, black) $p = \frac{1}{1000}$. The number of original flows is fixed to $N = 10^7$ and α is fixed to 1.5. The graphs are obtained with the numerical simulation described in Section 6.4.1.

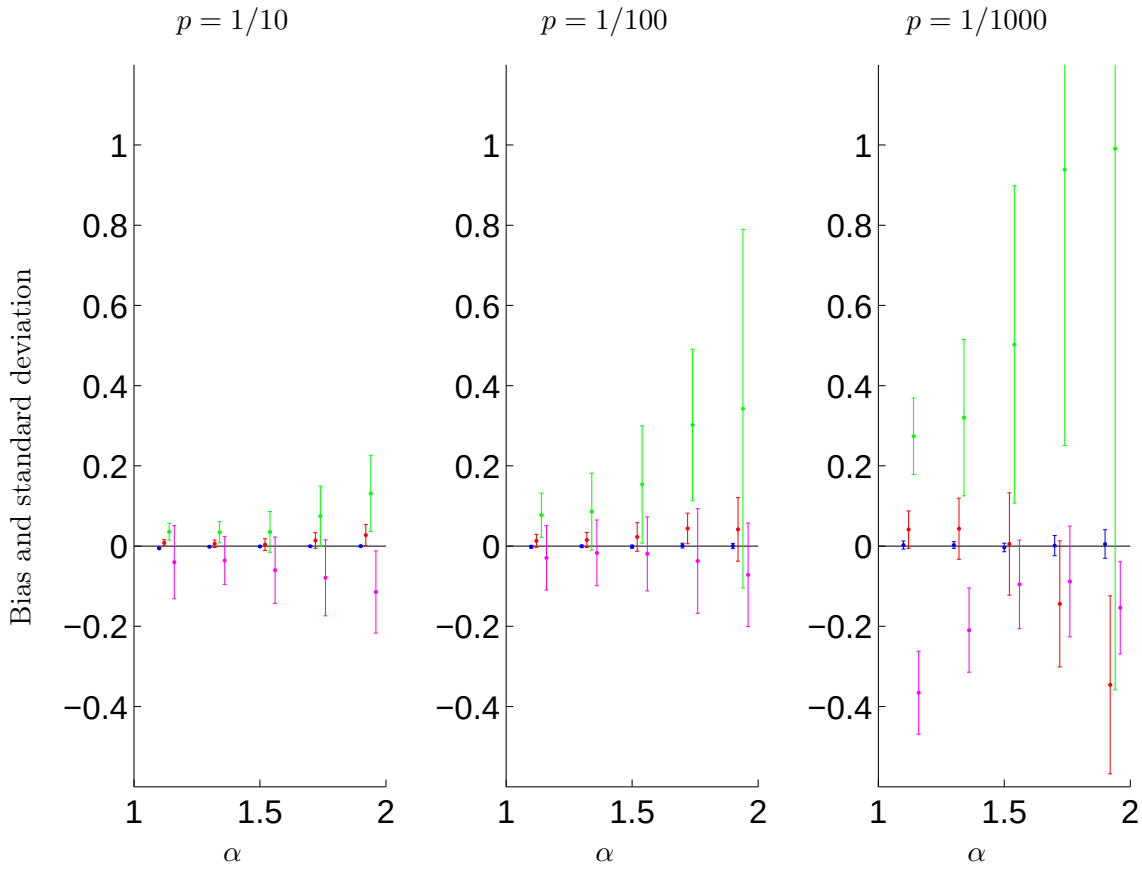

Figure 6.3: Comparison of the different tail-index estimation methods for five values of α: 1.1, 1.3, 1.5, 1.7, 1.9 (a small horizontal shift is introduced for clarity purposes): (blue) MLE – (red) Zeta *a priori* method – (green) scaling method – (magenta) stochastic-counting method. The error bars are centered on the bias and their lengths correspond to one standard deviation from the mean.

Comparison between the different estimators

Under the experimental conditions described above, Figure 6.3 displays the statistical performances of the different tail index estimators: the scaling method, the inverse approximation with a zeta *a priori*, the stochastic counting and the MLE. Every experimental result was obtained after we manually *optimized* each method's parameters, so that each of them reaches its minimum mean square error (MSE).

As a general remark, we stress that the performance of all the estimators systematically degrades with α. We put forward two causes to explain this. First, the estimation's difficulty is inherently increasing with as the scarcity of large flows grows (as α increases, the number of original elephants decreases and the number of large sampled flows also diminishes). Second, as we consider a fixed-length observation window, the number of original flows during this period grows as the tail distribution gets lighter (*i.e.*, α goes towards 2). These two competing phenomena should somehow balance. However, as we mentioned, $j_{\min} > 0$ is an important tuning parameter in all the methods, that forces to discard the small sampled flows. Given that its effect is more penalizing for large values

of α than for small ones, the number of observed flows effectively drops when α tends to 2, hence an increasing estimation variance.

Scaling method. This estimator shows a systematic bias which increases both with the tail index α and as the sampling rate p gets smaller. As we reported in [5], this poor performance certainly comes from the crude uniform *a priori* density choice and from the finite-support boxcar approximation. In practice then, the scaling method cannot be reliably used when $p \leq 1/10$.

Inverse approximation with zeta *a priori*. We had to select a sensible criterion to stop this iterative procedure: convergence is supposedly attained when the difference between two consecutive estimates of α becomes smaller than 0.005. In practice, this leads, in most cases, to a number of iterations between 10 to 100 iterations. As expected, compared to the scaling method, a more appropriate choice of the *a priori* law, along with a more adapted approximation of the binomial mass, sensibly reduces the estimation bias and variance. For $p \geq 1/100$, the bias even stabilizes with α. Still, for $p \leq 1/1000$, the influence of $j_{\min} > 1$ becomes too penalizing and the estimates turn rapidly unreliable for large values of α.

Stochastic counting method. The practical relevance of equation (6.18) depends on a correct choice of the threshold k_0. As for the other methods, this parameter was systematically tuned so as to minimize the MSE for each combination of the pair α and p. The observation period Δ is set to $\Delta = 5$ s. Compared to the previous approach, both bias and variance go up for almost all configurations. Yet, they remain remarkably steady as p goes from 1/10 to 1/100. Even more, the variance remains roughly constant with α and almost unchanged when the sampling rate falls to 1/1000. Its relatively poor performances notwithstanding, this striking stability is a valuable asset that prompts the use of stochastic counting with sparse thinning. Moreover, the extreme simplicity of the method allows for a responsive implementation at a very low computational cost.

MLE. For the sake of fairness, we discarded from this analysis the particular choice $j_{\min} = 0$, as it would involve non observable data, only retrievable from a deeper TCP packet inspection. Then, as the thorough MLE study of the next Section will show, MSE systematically increases with $j_{\min}$, and so we keep this index constant and equal to 1 in the following experiments.

It is clear from Figure 6.3, that MLE outperforms all the other methods. The variance of estimation is not only an order of magnitude below that of the inverse approximation with a zeta *a priori*, but also remains perfectly acceptable at very loose sampling rates. More precisely, we attain a precision up to the second decimal for $p = 1/100$, and up to the first decimal in the worse case corresponding to $p = 1/1000$ and $\alpha = 1.9$. Regarding the bias, the estimates of Figure 6.3 show no visible deviation from the theoretical values of α. The next Section reports on a more systematic evaluation of the MLE solution.

MLE performances

maximum-likelihood estimators are, by nature, asymptotically unbiased, as the numerical simulations of Figure 6.3 confirm. We undertook a complementary series of experiments to

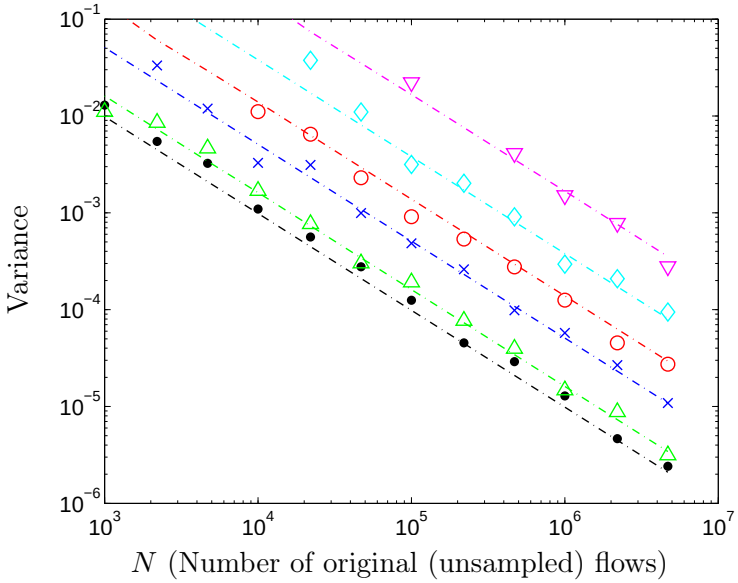

Figure 6.4: Variance of the MLE estimator against number of original flows considered for different values of N and $j_{\min}$. From the bottom to the top: ($\bullet$, black) $p = \frac{1}{10}$, $j_{\min} = 0$ – ($\triangle$, green) $p = \frac{1}{10}$, $j_{\min} = 1$ – ($\times$, blue) $p = \frac{1}{100}$, $j_{\min} = 0$ – ($\bigcirc$, red) $p = \frac{1}{100}$, $j_{\min} = 1$ – ($\Diamond$, cyan) $p = \frac{1}{1000}$, $j_{\min} = 0$ – ($\bigtriangledown$, magenta) $p = \frac{1}{1000}$, $j_{\min} = 1$. The dash-dotted plots represents the theoretical lower bound of the variance (Cramér-Rao bound). The value of α is fixed to 1.5

precisely evaluate the evolution of the vanishing bias when the number N of original flows grows to infinity, and for different sampling rates p. Fixing $\alpha = 1.5$, we then observed that beyond a number of original flows $N \geq 10^6$ the bias stays below 0.003 for all thinning cases ($p \geq 1/1000$) and for $j_{\min} = 0$. Under the same conditions though, it raises to 0.005 if, shifting $j_{\min}$ to 1, we discard the smallest sampled flows and reduce the number of effective observations. Numerical approximations needed to implement the MLE can also partially explain the residual bias.

We derived in Section 6.3.3 the Cramér-Rao bound associated to the estimation of the tail index α, from a sequence of n sampled flows. As n varies with $j_{\min}$, ($n = N$ if $j_{\min} = 0$, $n \leq N$ otherwise), it is empirically estimated for N, p and $j_{\min}$ fixed, and then used to evaluate the theoretical bound of inequality (6.27). For different values of p and $j_{\min}$, Figure 6.4 plots as a function of N, the empirical variances obtained from numerical simulations. Experimental points overlay almost perfectly with the theoretical limits, and prompt to the conclusion that the proposed MLE is efficient, even though we have no rigorous proof of this claim.

Now, analyzing the plots for N fixed, the variance of estimation naturally increases as the sampling process gets looser (*i.e.*, p gets smaller). It also increases with $j_{\min}$. This is clear from Figure 6.2 (c), which shows that the Fisher information is a decreasing function of $j_{\min}$ and thus the Cramér-Rao bound is an increasing function of $j_{\min}$. Let us recall here that the dominant effect in the deterioration of the variance when $j_{\min}$ increases is the consequent decrease of the number of observed flows (see Section 6.3.3 and Figure 6.2).

Compared to the variance amplitude, the bias of estimation is clearly negligible. Then,

regarding the mean square error (defined as the sum of squared bias and variance), it is primarily governed by the variance and behaves like $\mathcal{O}\left(N^{-1}\right)$. In particular, experimental results show that for a number of original flows larger than 10^6, the MSE does not exceed 10^{-4} when $p \geq \frac{1}{100}$, leading to a two decimal accuracy on the tail index estimate. Obviously, this outstanding precision, albeit very loose sampling, stems from a perfect match between data and the zeta distribution model. In the more general case though, where distributions are only asymptotically heavy-tailed, we are led to choose a larger value for $j_{\min}$. As a consequence, the effective number of sampled observed flows reduces accordingly and the performances of the MLE estimator can notably degrade.

Finally, let us stress that the computational cost of MLE, is an important drawback that can seriously hamper its use with real time constraints. However, storage of the important quantities for sampled values of α is a potential solution to overcome this limitation.

6.4.2 Confrontation to real traces

To evaluate the robustness of our maximum-likelihood estimation of α in the context of real traffic traces, we estimate α from an artificially sampled trace of internet traffic, captured on a real network link. We use the real trace acquired at *École Normale Supérieure de Lyon*, described in Section 4.3, where a `timeout` of 1 s is chosen for the flow reconstruction (but we check that other values yield almost the same results, so that this value is not critical).

The corresponding flow-size distribution, displayed in Figure 6.5 in log-log coordinates, clearly shows a heavy-tail behavior, and a zeta distribution model reasonably fits the data, provided that we discard flows of size smaller than some threshold $i_{\min}$. In our experiment, setting $i_{\min}$ to 35, we obtained a maximum-likelihood estimate of the tail index without sampling equal to $\hat{\alpha} = 0.9047$. We will use this value as a benchmark to compare the different sampling approaches.

To evaluate the accuracy of the different methods to estimate the flow size distribution tail index from sampled data, we artificially perform sampling on this trace with the three different sampling rates: $p = 1/10$, $p = 1/100$ and $p = 1/1000$. Again, flows of size smaller than some threshold $j_{\min}$ are discarded for the estimation of α. This threshold is set as follows: $j_{\min}$ is the smallest value of j verifying $B_p(i_{\min}, j) < \varepsilon$ where $j \geq pi_{\min}$ and ε is a threshold set to 0.05. Thus, every considered flow of size $j \geq j_{\min}$ is the sampled version of an original flow of size $i \geq i_{\min}$ with a probability greater than $(1 - \varepsilon)$. Practically, in our case, the values of $j_{\min}$ are $j_{\min} = 7$ for $p = 1/10$ and $j_{\min} = 2$ for $p = 1/100$ and $p = 1/1000$. For the estimations based on the stochastic counting method, the parameters k_0 and Δ are set to $k_0 = j_{\min}$ and $\Delta = 30$ s respectively.

Table 6.1 reports the estimated tail indices obtained with the different methods previously considered. Although all the methods give estimates of α roughly coherent with the expected value, two methods clearly stand out.

Firstly, the MLE clearly outperforms all the other methods in all the cases. This is a direct consequence of the maximum-likelihood principle which yields an adapted estimator, and it shows the relevance of the zeta distribution model utilization. Then, the inverse approximation method using a zeta *a priori* with a geometric mean approximation yields the closest estimates. This concordance is naturally interpreted by the fact that, as shown in Section 6.3.2, the zeta *a priori* with geometric mean approximation method shares the same basic estimation quantities ($\bar{\imath}_{(\alpha^{ap})}(j)$) as the MLE, and gives a good approximation

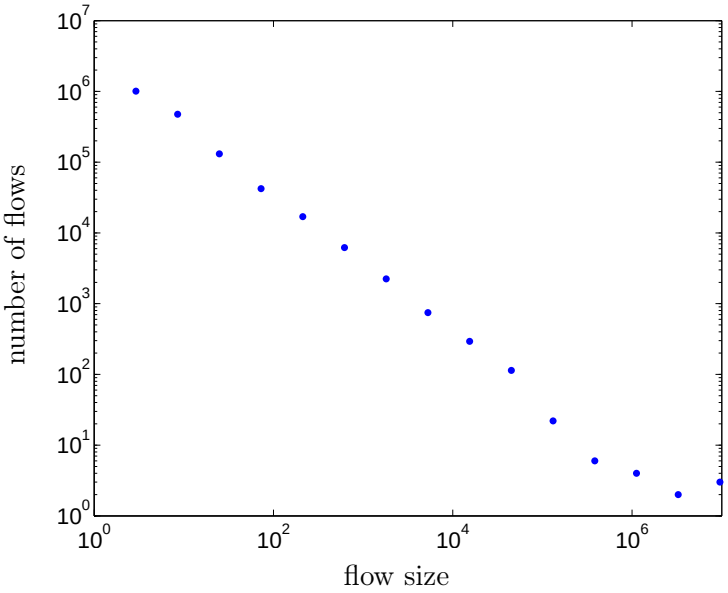

Figure 6.5: Flow-size distribution for the Internet traffic in a university context.

p	MLE	Zeta *a p.* with geom. mean approx.	Scaling	Stochastic counting
1	0.9047	-	-	-
1/10	0.9196	0.9281	0.9861	0.8413
1/100	0.9216	0.9935	1.2741	0.7050
1/1000	0.9572	1.0042	1.3160	1.0407

Table 6.1: Tail-index-parameter estimation for the Internet trace.

of the exact maximum-likelihood solution. Table 6.1 for this method reveals a small bias increasing as the sampling gets looser (p gets smaller), which is fully consistent with the results of Section 6.4.1 (Figure 6.3), for small values of α.

The last two methods give less accurate estimates. In adequacy with the results of Section 6.4.1 (Figure 6.3) for small values of α, the scaling method shows a positive bias, increasing as the sampling gets looser. This bias practically makes the use of the scaling method impossible for $p \leq 1/10$. The stochastic counting method seems to give reasonable estimates for any value of p. However, we want to stress here that this method relies on an appropriate choice of the observation period Δ, which has to be made with respect to a bias-variance tradeoff (too large a value of Δ deteriorates the estimation of $\mathbb{E}\{W_k\}$, thus increasing the variance, whereas too small a value of Δ might introduce an important bias). An inappropriate choice might make the estimation unstable and hamper practical use of this method.

Finally, we want to insist that the oustanding performance of the MLE comes at a high computational cost, whereas the stochastic counting method has a very low cost.

Conclusion

In this chapter, we tackled the question of estimating the flow-size distribution's tail index from packet sampled data. To address this question in a theoretically well-based framework, we used the zeta distribution as a paradigm of discrete heavy-tailed distributions.

We first reviewed the existing methods using simplifications, thus providing better insights into the nature of their approximations and highlighting the relationships between the different methods. In particular, we provided a theoretical justification for the (binomial) weighted geometric mean, a heavy-tailed data mapping that we had initially proposed in [5] in an ad-hoc manner. Then, to go beyond approximated solutions, we derived the exact maximum-likelihood estimation of the tail index. Our analytic solution was clearly shown to outperform other estimation variants and we reported very good results on simulated data and on a real traffic trace as an illustration.

For higher practical capabilities, further investigation is required to make our maximum-likelihood approach more efficient. Although theoretically sub-optimal some approximated methods are much faster in term of computational time and the possibility to find a good compromise between optimality and computational efficiency should be considered. Indeed, a computationally efficient estimation procedure using sampled data could prove very useful in the context of real-time adaptive protocols and network mechanisms, where the time constraint makes long computation based on the entire traffic observation impossible. Furthermore, the design of more robust estimators is also an interesting direction of research as maximum-likelihood approaches are likely to be very sensitive to data-model mismatches.

Finally, even if our ML estimator was developed within the specific context of network monitoring, it can readily apply to other situations of the same kind. For instance, this is the case with social networks where individuals (the equivalent of packets) are clustered into groups (the equivalent of flows) of heavy-tailed distributed sizes, and when only a cross-section of the population is observed.

Source-traffic characterization at flow scale.

- How to estimate the flow-size distribution's tail index from packet-sampled data?

We provide an exact maximum-likelihood solution for a zeta flow-size distribution, and show that most existing solutions are approximations of this MLE. Our MLE outperforms previously proposed solutions.

Chapter 7

Impact of heavy-tailed flow-size distributions on Quality of Service

This chapter describes an early-stage research, and the preliminary empirical results presented here have not yet been published.

Impact of the traffic characteristics on QoS (in performance terms).

- In realistic conditions implying finite buffer sizes, what is the impact of heavy-tailed flow-size distributions on QoS (in performance terms)? More precisely, on which QoS metric?

Introduction

In the last two chapters, we have seen how heavy-tailed flow-size distributions generate long-range dependence in the aggregate traffic, and how to estimate the tail index α_{SI} of these distributions from sampled packet data. In this chapter, we now tackle the question of the impact of the parameter α_{SI} on the Quality of Service, in performance terms.

We have seen in Section 3.2 of Chapter 3 that many theoretical results appeared recently on the impact of α_{SI} on QoS. However, most of them give asymptotic results, either for very large or very small buffers (see [Mandjes and Boots, 2001]), and do not consider the impact of the TCP feedback. These results predict a degradation of QoS with small values of α_{SI} (*i.e.* very heavy tails) when the buffer is large, and an insensitivity of QoS with respect to α_{SI} for small buffers. A numerical evaluation (based on ns2 simulations) of the impact of α_{SI} with the TCP protocol appeared in [Park et al., 1997], where the authors reach the same conclusions. However, more recent studies [Arvidsson and Karlsson, 1999, Ben Fredj et al., 2001] argued that only the mean of the flow-size distribution is important. In this chapter, our goal is to provide qualitative insights into the impact of the value of the tail index α_{SI} in practical situations implying finite-size buffers of medium size, for the two protocols UDP and TCP.

The tail index α_{SI} describes the algebraic decay of the distribution in the asymptotic regime of very large flows, but does not depend on the body of the distribution. However, the Pareto distribution has only two parameters: μ_{SI} and α_{SI}. The mean flow time μ_{SI} is an important parameter, which controls the small-scale correlation of the traffic and the number of active flows. It has been posited as a major control parameter [Arvidsson and

Karlsson, 1999,Ben Fredj et al., 2001], which is why we keep it constant in our simulations. Thus, varying the α_{SI} parameter not only changes the tail of the distribution, but actually modifies the entire distribution. To assess the role of the distribution's tail parameter α_{SI}, we will systematically compare the results obtained using two different Pareto-like distributions whose parameters μ_{SI} and α_{SI} are the same, but whose bodies are different:

$$(P1) \qquad p_1(s) = \frac{\alpha_{\mathrm{SI}} k_1^{\alpha_{\mathrm{SI}}}}{(s+k_1)^{(\alpha_{\mathrm{SI}}+1)}},\ s > 0, \tag{7.1}$$

$$(P2) \qquad p_2(s) = \frac{\alpha_{\mathrm{SI}} k_2^{\alpha_{\mathrm{SI}}}}{s^{(\alpha_{\mathrm{SI}}+1)}},\ s > k_2, \tag{7.2}$$

where $k_1 = (\alpha_{\mathrm{SI}} - 1)\mu_{\mathrm{SI}}$ and $k_2 = \frac{(\alpha_{\mathrm{SI}}-1)\mu_{\mathrm{SI}}}{\alpha_{\mathrm{SI}}}$. Note that the distribution (P1) is the distribution introduced in equation (2.9) (in Section 2.2 of Chapter 2) that we have used in previous chapters.

As we have mentioned in Chapter 1, questions related to QoS reveal different aspects if the protocol used is UDP or TCP (because of the feedback introduced by TCP), and we used different approaches to deal with these two cases.

7.1 The UDP case: numerical study

In the case of UDP traffic, the source emission process is independent of the system's dynamic and we use matlab simulations described in Section 4.4 of Chapter 4. The typical meaningful QoS metrics for UDP traffic that we consider here are loss rate and buffer occupancy.

7.1.1 Simulations' description

Most of our simulations consist of simulated traffic from 50 ON/OFF sources, whose parameters are summarized in Table 7.1. The aggregate traffic is averaged in basic time windows of size $\Delta_0 = 0.2$ ms and used as the input of our continuous queue simulator. Our queueing simulations cover a large spectrum of buffer sizes B between 20 and 10^5 pkts. As a meaningful quantity, we define quantity $\Delta_B = \frac{B \times pktsize}{C}$, which represents the buffer size in time units, or the *buffer scale*. We verified that for buffers of size Δ_B larger than Δ_0, continuous and discrete queue simulations give the same results, confirming the ability of the buffer to absorb traffic variations at scales smaller than Δ_B.

Using continuous queue simulations allows us to flexibly manipulate the input traffic. Then, we can impose a load of exactly 0.95 by adjusting the mean of the aggregate traffic. To assess the role of multiplexing on the queueing behavior, we also easily simulate the aggregate traffic from 200 independent sources by dividing the variance of the aggregate traffic from 50 independent sources by 4.

7.1.2 Results and discussion

Loss rate

Impact of α_{SI}. Figure 7.1 (top) displays the loss rate as a function of the buffer size Δ_B. For each distribution (P1) or (P2), comparison of the graphs for the different values of α_{SI} clearly brings the same conclusion: the loss rate is independent of α_{SI} for small buffers,

Parameter	value
Bottleneck	$C = 1$ Gbps
Buffer size	$20 - 10^5$ pkts
Source nb.	50
Source rate	38 Mbps
load	0.95
Experiment duration	$T = 2$ hours
Flow-size distribution	heavy-tailed, (P1) or (P2)
	$\alpha_{\rm SI} = 1.1, 1.5, 1.9$ or 3.0
Mean ON time	$\mu_{\rm ON} = 0.0316$ s
OFF time distribution	exponential
Mean OFF time	$\mu_{\rm OFF} = \mu_{\rm ON}$

Table 7.1: Simulations' parameters

and decreases when $\alpha_{\rm SI}$ increases for large buffers (the heavier tail, the higher loss rate). This result is fully coherent with the asymptotic results of [Mandjes and Boots, 2001]. To specify the bound between those two regimes, Figure 7.1 (bottom) displays the standard deviation of the aggregate traffic averaged at scale Δ_B. As we have already commented in Chapter 5, the standard deviation depends on $\alpha_{\rm SI}$ only beyond the *knee* $\mu_{\rm ON}$. Interestingly, we observe that the critical value of Δ_B beyond which the loss rate depends on $\alpha_{\rm SI}$ does not correspond to this same *knee* (it is actually much smaller). This indicates that the standard deviation of the input traffic at scale Δ_B is not sufficient to evaluate the loss rate with a buffer of size Δ_B. Instead, correlations at scales larger than Δ_B play an important role. This result is fully coherent with the critical time-scale approximation and the subsequent approximation of [Ribeiro et al., 2006b] (see Section 3.2 of Chapter 3), showing that the loss rate (or more precisely in these theoretical works, the overflow probability) depends on the marginal distribution of the input traffic at all time scales. Practically, here, the loss rate significantly depends on $\alpha_{\rm SI}$ even for buffers of size much smaller than $\mu_{\rm ON}$, for both distributions (P1) and (P2).

Impact of the distribution shape. We now turn to the comparison of the results obtained with distributions (P1) and (P2) for a same value of $\alpha_{\rm SI}$. Figure 7.1 (top) shows that for each value of $\alpha_{\rm SI}$, distribution (P1) leads to higher loss rates than (P2). This holds for values of Δ_B larger than a threshold approximately equal to the critical buffer size mentioned above, beyond which the loss rate depends on $\alpha_{\rm SI}$. Figure 7.1 (bottom) on the other side shows that the standard deviation of the input traffic is larger with distribution (P1) only above the classical knee $\mu_{\rm ON}$. Based on the results discussed above, it is not surprising that this differentiation of the standard deviations beyond $\mu_{\rm ON}$ has an effect on the loss rate for much smaller buffers. To interpret the variations of the standard deviation, Figure 7.2 displays distributions (P1) and (P2) for each value of $\alpha_{\rm SI}$ and the corresponding log-diagrams of the aggregate traffic. This figure shows that while the algebraic decay of the distribution is the same with (P1) and (P2), the asymptotic regime is achieved for larger flow sizes with distribution (P1), resulting in a higher number of large flows. This effect increases when $\alpha_{\rm SI}$ increases and the ratio between the two

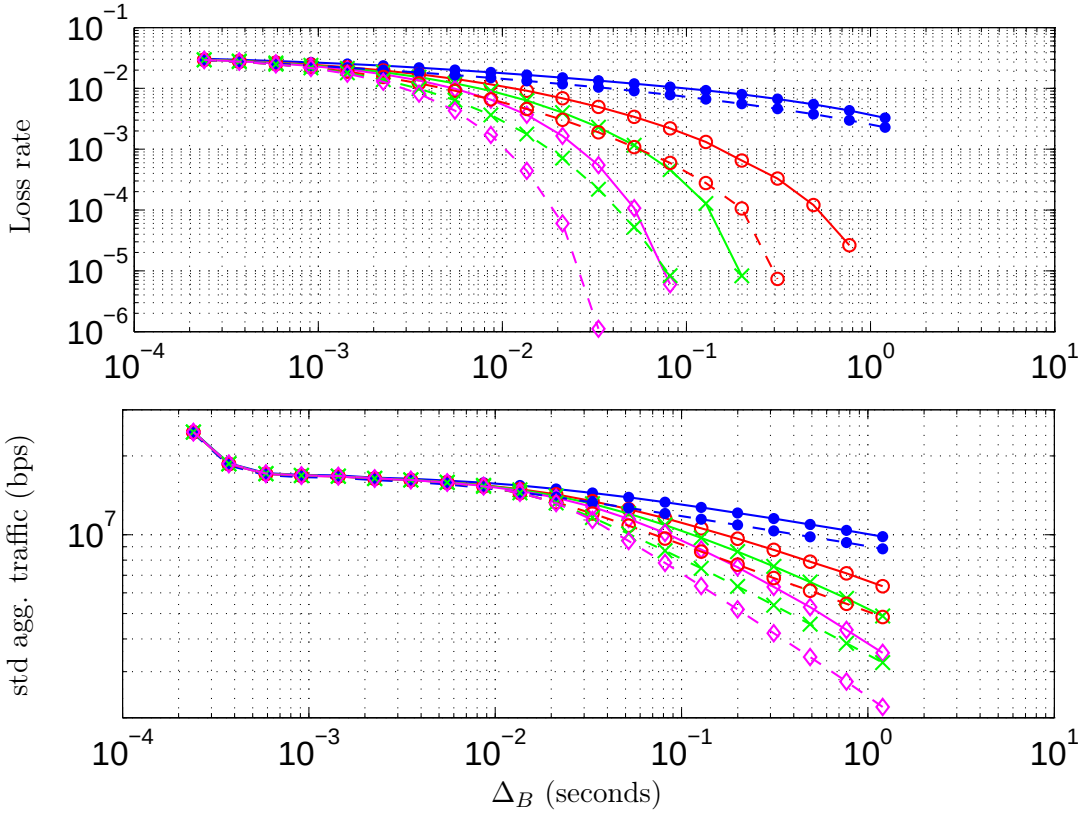

Figure 7.1: Loss rate and standard deviation of the aggregate traffic versus Δ_B for different values of $\alpha_{\rm SI}$ and for distributions (P1) and (P2). (continuous lines) The flow-size distribution follows the law (P1) – (dashed lines) The flow-size distribution follows the law (P2). ($\bullet$, blue) $\alpha_{\rm SI} = 1.1$ – ($\circ$, red) $\alpha_{\rm SI} = 1.5$ – ($\times$, green) $\alpha_{\rm SI} = 1.9$ – ($\diamond$, magenta) $\alpha_{\rm SI} = 3.0$. The missing points on the right part of the loss-rate curves for large values of $\alpha_{\rm SI}$ correspond to situations where no loss was observed in the simulation.

distribution tails can be simply related to the ratio of the pre-factors of distributions (P1) and (P2) for large flow sizes: $\alpha_{\mathrm{SI}}^{\alpha_{\mathrm{SI}}}$ (see equation (7.2)). Similarly, the asymptotic regime is also achieved at a larger scale in the log-diagrams with distribution (P1), and corresponds to a higher absolute value of the variance at large scales. As compared to distribution (P2) for the same mean and tail index, the use of distribution (P1) then leads to a larger loss rate for large buffers.

Impact of multiplexing. To conclude our discussion on the impact of α_{SI} on the loss rate, we comment on the effect of multiplexing. Since we have seen that for the two distributions (P1) and (P2) the impact of α_{SI} is the same, we discuss here only the results for distribution (P1) displayed on Figure 7.3. The input traffic corresponds to the aggregation of 200 sources yielding the same load of 0.95. Although the loss rates are smaller than on Figure 7.1 (top) which corresponds to a traffic of 50 sources (because the standard deviation of the input traffic is smaller), the same variations with α_{SI} are observed. Consequently, the same conclusions hold: the loss rate depends on α_{SI} only beyond some threshold, significantly smaller than μ_{ON}. We conclude that, as pointed out in several papers (see Section 3.2 of Chapter 3), multiplexing only affects a pre-factor of the loss rate, and does not modify the impact of α_{SI}.

Buffer occupancy

We now turn to the buffer occupancy. To comment on the impact of α_{SI}, we concentrate only on the case where the input traffic corresponds to the aggregation of 50 sources, whose flow-size distribution follows law (P1). Figure 7.4 shows the mean buffer occupancy as a function of Δ_B. Clearly, the same conclusion as for the loss rate holds: the mean buffer occupancy is insensitive to α_{SI} for small buffers, and decreases when α_{SI} increases for large buffers (the heavier the tail, the larger the mean buffer occupancy). However, contrarily to the case of the loss rate, the knee μ_{ON} is a rather accurate critical value to distinguish those two regimes. We propose the following interpretation of this result. Given a buffer scale Δ_B, the correlations of the aggregate traffic at larger scales tend to increase the loss rate because they induce long periods where the aggregate traffic is larger than the capacity, thus feeding the buffer beyond its size. However, during these periods, the buffer occupancy saturates at a value corresponding to its size. These long periods where the buffer occupancy is equal to its size are also compensated by long periods where the aggregate traffic is smaller than the capacity and the buffer occupancy is zero. This compensation leads to a mean buffer occupancy that does not depend on the correlations at scales larger than Δ_B, but instead depends mainly on the variance at scale Δ_B. In other words, the mean buffer occupancy is a first order statistic and depends mostly on the (mean and) variance of the input traffic at scale Δ_B; whereas the loss rate is a more complex property of the buffer occupancy distribution that depends also on the correlations at larger scales.

To complete the description of the buffer occupancy, Figure 7.5 displays the buffer-occupancy distributions for 3 different buffer sizes (small, medium and large). This figure shows a complex behavior of the occupancy distribution, but allows us to make a few general comments. Firstly, as it was the case for the mean delay, the distribution significantly depends on α_{SI} only for large buffers. Also, the distribution of the delay for small buffers is not a truncated version of the distribution for large buffers. This is due to the

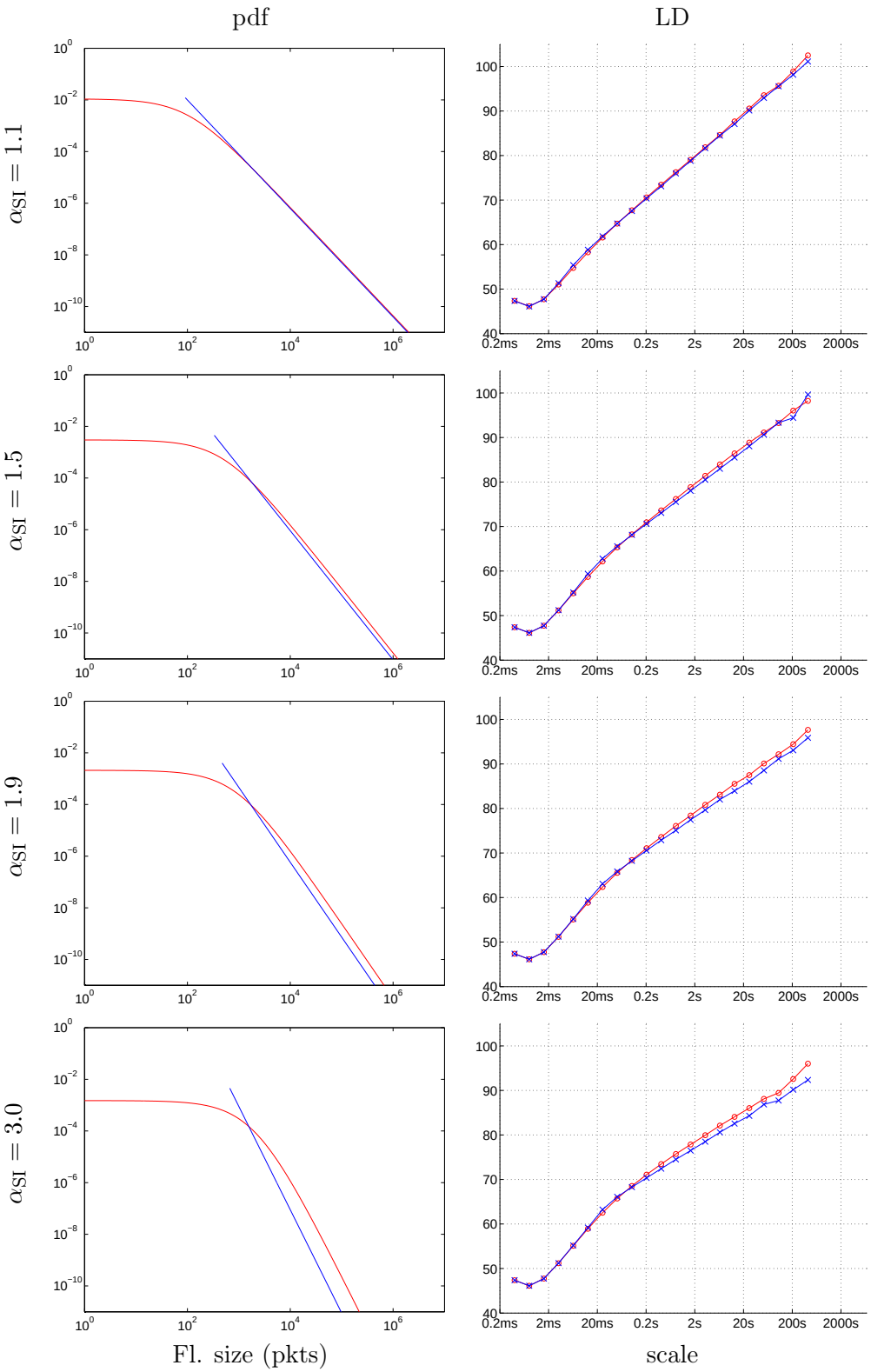

Figure 7.2: Flow-size distributions and corresponding log-diagrams of the aggregate traffic of 50 sources. (red) (P1) – (blue) (P2). The LDs have not been vertically shifted and are therefore presented with their absolute position. The distributions displayed are the theoretical pdfs of equation (7.2)

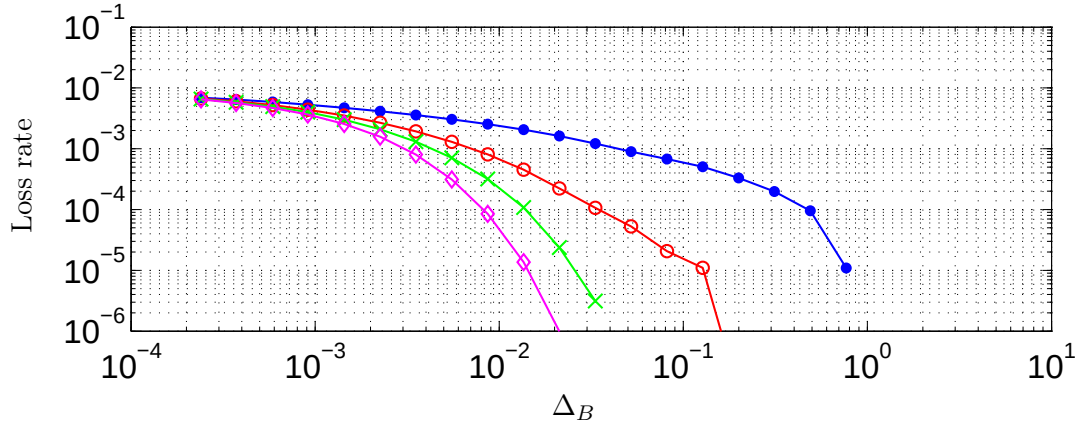

Figure 7.3: Loss rate versus Δ_B, for an aggregate input traffic from 200 sources with load 0.95. (•, blue) $\alpha_{\mathrm{SI}} = 1.1$ – (∘, red) $\alpha_{\mathrm{SI}} = 1.5$ – (×, green) $\alpha_{\mathrm{SI}} = 1.9$ – (⋄, magenta) $\alpha_{\mathrm{SI}} = 3.0$

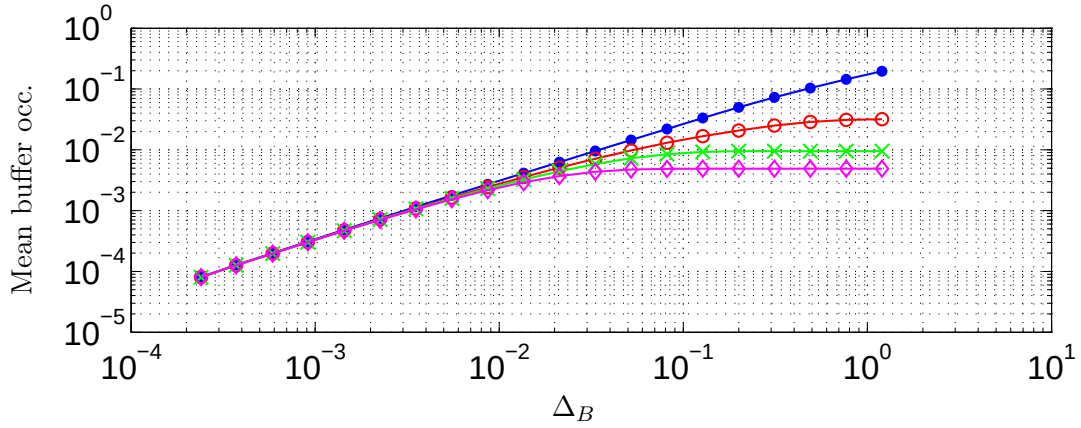

Figure 7.4: Mean buffer occupancy as a function of Δ_B. (•, blue) $\alpha_{\mathrm{SI}} = 1.1$ – (∘, red) $\alpha_{\mathrm{SI}} = 1.5$ – (×, green) $\alpha_{\mathrm{SI}} = 1.9$ – (⋄, magenta) $\alpha_{\mathrm{SI}} = 3.0$. The constant parts observed for large values of α_{SI} and large buffers correspond to loss-free situations where increasing the buffer size does not modify its dynamic.

saturation effect, characteristic of finite-size buffers that we discussed above. It illustrates the difference between the loss rate in a finite-size buffer and the overflow probability in an infinite-size buffer, mentioned in Section 3.2 of Chapter 3, which although following similar trends, are different quantities.

7.1.3 Conclusion

In this section, we showed that for UDP traffic, the QoS metrics loss rate and buffer occupancy are degraded with very heavy tails (small values of the tail index α_{SI}) if the buffer is "large enough". We showed that this result holds independently of the distribution used ((P1) or (P2)), even if the chosen distribution has an effect on the absolute values of these QoS quantities. For the mean buffer occupancy, we showed that the classical *knee* μ_{ON} is a good critical value to distinguish small and large buffers. On the opposite, the loss rate is sensitive to α_{SI} even for much smaller buffers. We interpreted this difference by the fact that the mean buffer occupancy, as a first-order statistic, is mostly sensitive to the aggregate input traffic characteristics at scale Δ_B, whereas the loss rate, as a more complex property of the distribution, is also sensitive to the correlations at larger scales. A more comprehensive set of experiments would be required to further investigate this preliminary interpretation. In particular, we used a Gaussian input traffic which is then characterized only by its variance at each scale. Given the non-linear effects of the queueing behavior, it is likely that the situation would be more complex in a more general context. This will be the object of future work.

7.2 The TCP case: experimental study

In the case where the TCP protocol is used, the analysis of QoS is more complex and a simple simulation-based approach as we used for UDP is not possible. The feedback mechanism makes impossible an open-loop analysis, where the traffic is first generated, and then given as the input of a queue simulator. Moreover, the load is not controlled. Instead, it is a consequence of the TCP mechanisms adapting the emission rate of each source. The sources are also no longer independent: their interaction at the buffer introduces correlations at the RTT scale. Finally, as we have seen in Section 5.2.2 of Chapter 5, the aggregate traffic is strongly non-Gaussian and thus cannot be characterized simply by its variance. For these reasons, the question of the impact of heavy-tailed flow-size distributions is more delicate in the case of TCP than in the case of UDP, and the conclusions can be different. Among the answers proposed in the literature accounting for the TCP feedback (see Section 3.2 of Chapter 3), different conclusions have been reached. In [Park et al., 1997], the authors conclude, based on ns2 simulations, that a very heavy tail (*i.e.*, a small tail index) systematically degrades the performance. On the other hand, in [Arvidsson and Karlsson, 1999, Ben Fredj et al., 2001], the authors propose models showing that first-order performance statistics (*e.g.*, mean throughput and flow completion times) depend only on the mean of the flow-size distribution.

In this section, we provide results of experiments performed on our metrology platform described in Section 4.2 of Chapter 4, attempting to provide qualitative insights into the impact of α_{SI} on QoS. Our approach is similar to previous section: we systematically compare results using the two distributions (P1) and (P2).

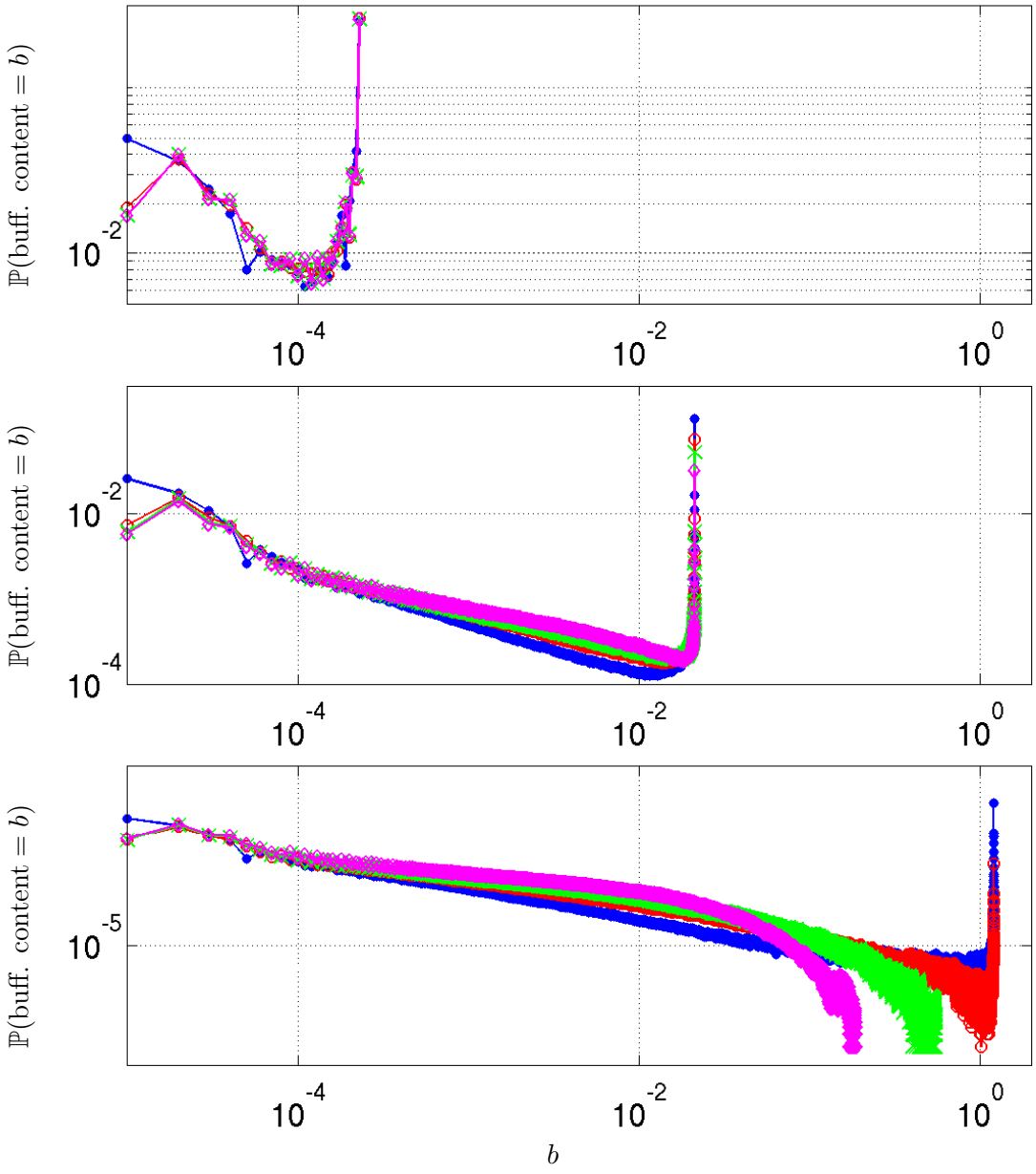

Figure 7.5: Delay distribution for 3 different buffer sizes: (top) $B = 20$ pkts, $\Delta_B = 0.24$ ms – (middle) $B = 1769$ pkts, $\Delta_B = 21.2$ ms – (bottom) $B = 10^5$ pkts, $\Delta_B = 1.2$ s. ($\bullet$, blue) $\alpha_{\mathrm{SI}} = 1.1$ – ($\circ$, red) $\alpha_{\mathrm{SI}} = 1.5$ – ($\times$, green) $\alpha_{\mathrm{SI}} = 1.9$ – ($\diamond$, magenta) $\alpha_{\mathrm{SI}} = 3.0$

Parameter	value
Topology	Figure 4.3
Other site	Nancy
Minimal RTT	10 ms
Bottleneck	$C = 1$ Gbps
Buffer size	96, 896 pkts
Source nb.	45
Experiment's duration	$T = 2$ hours
Flow-size distribution	(P1), (P2)
	$\alpha_{\mathrm{SI}} = 1.1, 1.5, 1.9$ or 3.0
Mean flow size	$\mu_{\mathrm{SI}} = 1000$ pkts
Mean ON time	experiment-dependent
	around 0.3 s
OFF time distribution	exponential
Mean OFF time	0.3 s

Table 7.2: Experiments' parameters

7.2.1 Experiments' description

All our experiments consist in 2 hours of traffic generated from 45 sources, from which we discard the first 5 minutes to avoid transient behavior at the beginning of the experiment. To avoid interactions of the sources unrelated to the buffer dynamic, we use only one source per node. Table 7.2 summarizes the important experiments' parameters. As we did in the previous section, to assess the role of the flow-size distribution's tail index, we impose flow-size distributions of various tail indices, and systematically compare the results with the two types of Pareto-like distributions (P1) and (P2). The mean flow size is always set to $\mu_{\mathrm{SI}} = 1000$ pkts. To ensure that this mean flow size is exactly respected (even for $\alpha_{\mathrm{SI}} = 1.1$), we artificially multiply the generated random series corresponding to the flow sizes by an appropriate factor. Since the TCP protocol adapts the source rate, the ON times are not imposed. Instead, they depend on the experimental conditions and become an interesting QoS metric. In all of our experiments, however, we observed that the mean ON time is always around 0.3 s, and has the same value as the imposed mean OFF time.

Each experiment is performed two times with two different buffer sizes: 96 pkts and 896 pkts, corresponding to values of Δ_B of respectively 1.2 ms and 10.8 ms. These two values are smaller than the *knee* μ_{ON}. However, in the case of TCP, it is not realistic (and in most cases not possible) to have a buffer size larger than μ_{ON}. Indeed, since the packets are emitted by bursts at each RTT, and the RTT increases with the buffer occupancy, which increases with the buffer size, this situation inevitably leads to a mean flow duration larger than the buffer size.

During the experiments, the aggregate output is captured as well as the input for one of the sources. The results presented for the loss and delay are deduced from the comparison of the input and output captures of this source. For the case of losses, we verified that the loss rates obtained are coherent with *netstat* data collected at each source.

7.2.2 Results and discussion

We now present and discuss the results of our experiments in term of QoS metrics of two types: buffer QoS metrics (loss, packet delay and throughput) and end-to-end QoS metrics. For the latter, we choose to focus on the flow rate, which is the inverse of the completion time.

Buffer QoS: loss, delay, throughput.

Figure 7.6 displays the evolution of the buffer QoS metrics with α_{SI}, for the two distributions (P1) and (P2) and the two buffer sizes. It shows that these evolutions are less clear in the case of TCP than they were in the case of UDP: the variations with α_{SI} are within a rather small range, and each point is associated with a rather large error bar, due to the natural variability of real experimental conditions and to the complexity of a real system. However, some tendencies and general conclusions can be drawn on the impact of α_{SI}. Firstly, as a general remark, we observe that for $\alpha_{\mathrm{SI}} = 1.1$, the results corresponding to the two distributions (P1) and (P2) are almost undistinguishable, because the distributions themselves are very close (see Figure 7.2). When α_{SI} increases, the difference between distributions (P1) and (P2) increases, and the deviation between the corresponding QoS curves of Figure 7.6 also increases in most cases. Regarding the loss rate, we observe that for distribution (P1) (red curves), it is constant, whereas for distribution (P2), it decreases when α_{SI} increases. It indicates that the degradation of the loss rate with very heavy tails observed in [Park et al., 1997] is not a universal behavior, valid for all experimental conditions (in their experiments, the loss rate was much higher, of the order of 4%). It also shows that for the same experimental conditions, two heavy-tailed flow-size distributions with the same mean can lead to different behaviors of the loss rate as a function of α_{SI}. It indicates that the parameter α_{SI} in itself is not sufficient to evaluate the loss rate, and is not sole responsible for its evolution.

For the two other metrics, mean packet delay and throughput, the evolution with α_{SI} is the same for the two distributions (P1) and (P2). The mean packet delay systematically decreases when α_{SI} increases, indicating a performance degradation for very heavy tails (small values of α_{SI}). On the opposite, the decrease of the mean throughput when α_{SI} increases indicates a performance improvement for very heavy tails. Note that the throughput is calculated from the capture acquired after the buffer and thus does not take lost packets into account. However, given the small loss rates, the throughput of the input traffic and output traffic should not be significantly different. While the evolution of the mean packet delay is coherent with the trend observed in [Park et al., 1997], the evolution of the throughput is in contradiction with both the observations of [Park et al., 1997], where link utilizations increase when α_{SI} increases, and the prediction of the model of [Ben Fredj et al., 2001], that throughput depends only on the mean of the distribution. However, as we shall now develop, the tendencies that we observe in our experiments are mainly governed by long transient behaviors at the beginning of each flow.

End-to-end QoS: flow rate

Figure 7.7 displays the distribution of the mean flow rates as a function of the flow sizes for each experiment, and the corresponding flow-size distributions, where flow sequences are reconstructed with a `timeout` of 100 ms from the output capture.

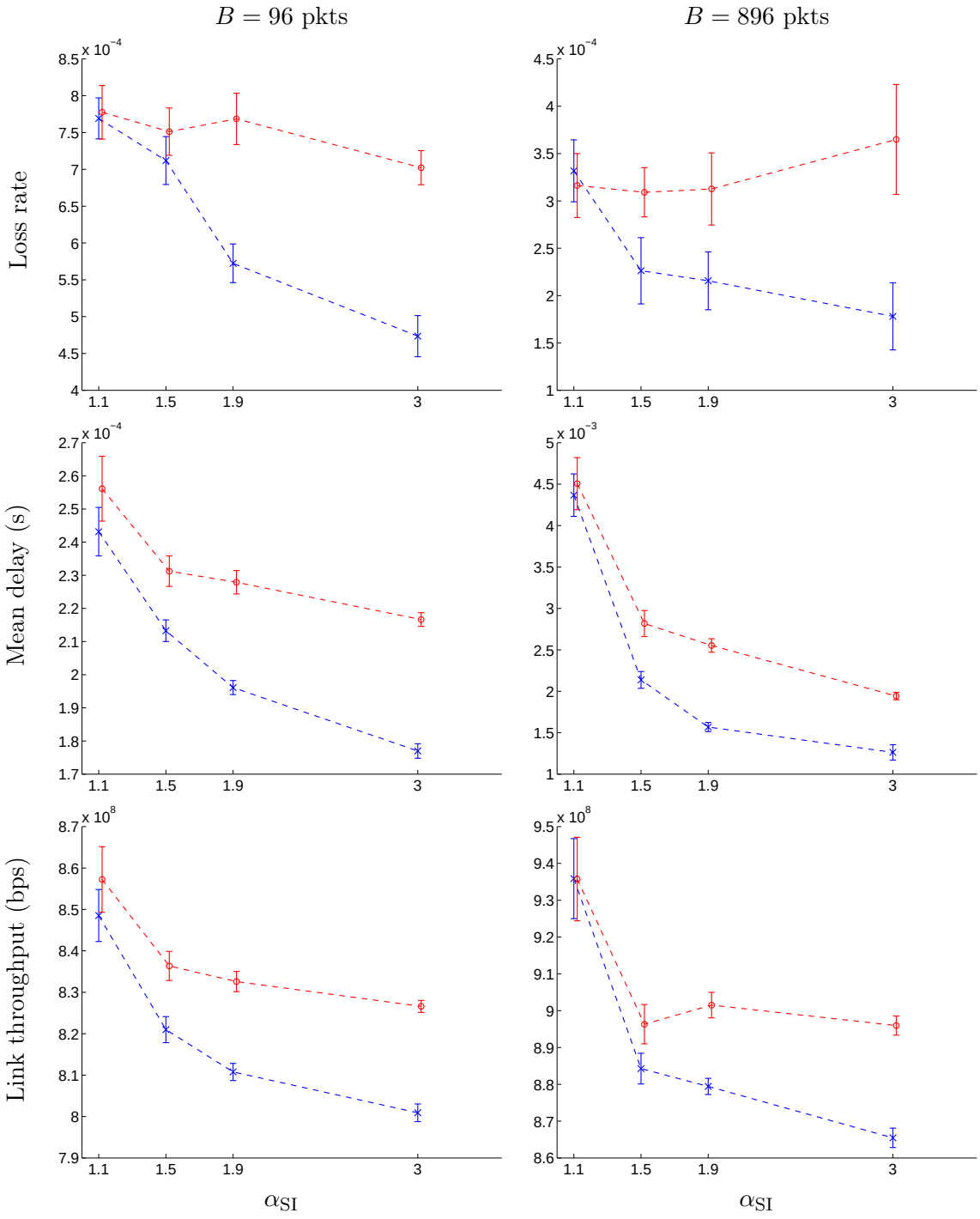

Figure 7.6: Loss rate as a function of α_{SI}. ($\circ$, red) flow-size distribution (P1) – ($\times$, blue) flow-size distribution (P2). The error bars are estimated by a bootstrap method [Efron and Tibshirani, 1993] using 25 sub-time windows which we observed almost independent.

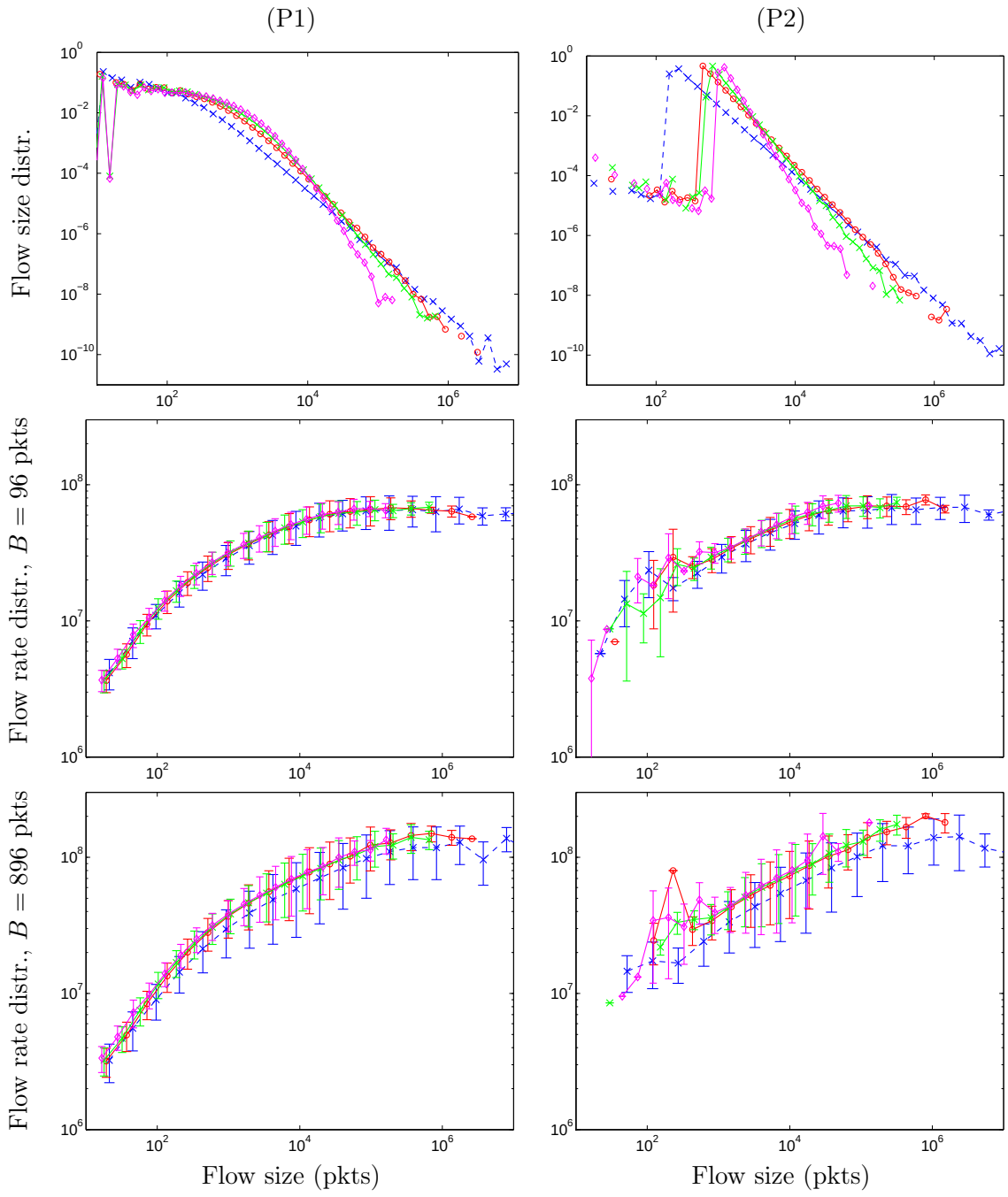

Figure 7.7: Flow-size and rate distributions, where flows are reconstructed with a `timeout` of 100 ms from the output capture: (left) for distribution (P1) – (right) for distribution (P2). (top) Flow-size distribution, from the experiments with a buffer $B = 96$ pkts – (middle, bottom) Mean flow-rate distributions in logarithmically spaced bins for the flow sizes. The error bars corresponds to one standard deviation from the mean in each bin, and are dissymmetrical because of the logarithmic scale. (•, blue) $\alpha_{\mathrm{SI}} = 1.1$ – (○, red) $\alpha_{\mathrm{SI}} = 1.5$ – (×, green) $\alpha_{\mathrm{SI}} = 1.9$ – (◇, magenta) $\alpha_{\mathrm{SI}} = 3.0$.

The estimated flow-size distributions corresponds to the experiments with a buffer of size $B = 96$ pkts. While they exhibit a global shape closely matching the theoretical distributions (see Figure 7.2), a small number of small flows are also present, especially noticeable for distribution (P2). They correspond to larger flows split by the timeout mechanism, and are naturally less numerous when the buffer size is larger ($B = 896$ pkts) and the loss rate smaller.

Regarding the flow rate distributions, the most striking observation is that they are increasing until some rather large threshold ($10^4 - 10^5$ pkts for $B = 96$ pkts and $10^5 - 10^6$ pkts for $B = 896$ pkts) before stabilizing for large flow sizes. To explain the origin of this behavior, Figure 7.8 displays the evolution of the rate of one typical large flow taken from the experiment with (P1), $B = 96$ pkts and $\alpha_{\mathrm{SI}} = 1.1$. The top figure shows that the mean rate of this flow increases as the number of transmitted packets increases, and stabilizes after a threshold roughly equal to the threshold of Figure 7.7 (left, middle) corresponding to the same experimental conditions. Figure 7.8 (bottom) shows that this is due to the long transient behavior at the beginning of the flow. Roughly, the first loss occurs after 2 seconds and has an effect on the mean flow rate until around 5 seconds (corresponding to the transmission of 28,000 packets). This long transient behavior is due to the absence of slow start at the beginning of the flow. Indeed, as shown on Figure 7.8 (bottom), the rate increases linearly, from the beginning (the different slopes correspond to different RTTs due to variations of the queueing delay). The end of the slow-start phase is controlled by an optimization parameter of the TCP implementation called `SSTRHESHOLD`. In our Linux configuration, this parameter can only decrease from one flow to the next one (of the same machine). Due to the high load in our experiments, it rapidly stabilizes at around 12 packets, so that we almost only observe the congestion-avoidance phase. This decreasing behavior of the `SSTRHESHOLD` is also due to the fact that in our experiments, the destinations are always at the same location and the connections always have similar RTTs. This can be for example representative of the particular case of a distributed application which collects data always on the same server. The high loads of our experiments can also be representative of an access point of the Internet.

The observation of a larger number of individual flows yields the same conclusion and proves that the behavior of the mean flow rate as a function of the flow size (Figure 7.7) is due to the long transient behaviors at the beginning of each flow. Smaller loss rates observed with a buffer size of $B = 896$ pkts induce even longer transient periods, which explains that the threshold is higher on Figure 7.7 (bottom, $B = 896$ pkts) than on Figure 7.7 (bottom, $B = 96$ pkts). Comparing the flow-rate distributions of Figure 7.7 for the two buffer sizes, we also see that variances are increased for the larger buffer size, due to the higher variability of the queueing delay. Finally, except for the experiment with $\alpha_{\mathrm{SI}} = 1.1$ and $B = 896$ pkts where the flow rates are slightly smaller (most probably due to an higher mean delay increasing the RTT), the flow-rate distributions do not exhibit significant variations for the different values of α_{SI}, independently of the distribution used ((P1) or (P2)) and of the buffer size.

We now come back to the results of the buffer QoS, that we partially interpret using our observations on the flow-rate distributions. The mean link throughput of Figure 7.6 (bottom) can be obtained (up to some constant depending on the mean number of active flows which is almost constant in all of our experiments) as the mean of the flow-rate distribution of Figure 7.7 (bottom, middle) weighted by the flow-size distribution of Figure 7.7 (top). For a given distribution ((P1) or (P2)), the mean throughput is then larger for

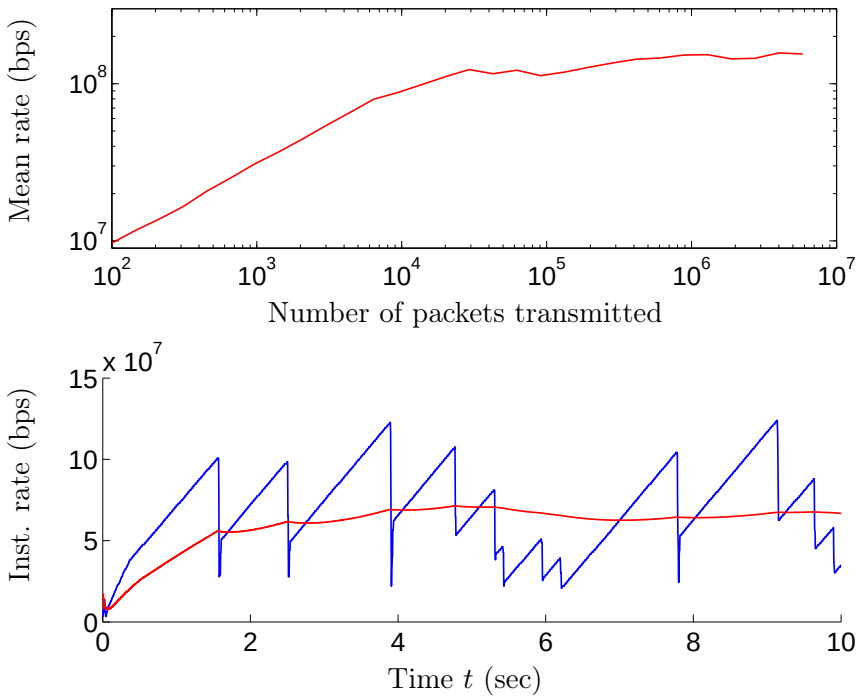

Figure 7.8: Analysis of one large flow of the experiment with (P1), $B = 96$ pkts and $\alpha_{\rm SI} = 1.1$. (top) Mean rate of the flow as a function of the number of transmitted packets – (bottom) In blue: evolution of the instantaneous rate of the flow accross time zoomed on the beginning of the flow (the first 10 seconds). The red curve corresponds to the mean flow rate between time 0 and t. These graphs are obtained from web100 data collected every 3 ms.

small values of $\alpha_{\rm SI}$, because of a larger number of large flows achieving high rates. For a given value of $\alpha_{\rm SI}$, it also explains the larger throughput obtained with distribution (P1) because the number of large flows is greater (see Figure 7.2). It shows that, more than the value of $\alpha_{\rm SI}$ reflecting the algebraic decay of the distribution for large flow sizes, the absolute values of the whole distribution are essential to determine the link throughput. The interpretation of the mean delay and loss rate is more delicate, because it depends not only on the mean input traffic, but also on its distribution, variability and correlations. The load decrease observed in most cases when $\alpha_{\rm SI}$ increases partly explains the decrease of the mean delay. Regarding the evolution of the loss rate (in particular in the situation of a large buffer ($B = 896$ pkts) with distribution (P1) where the mean delay decreases whereas the loss rate remains constant when $\alpha_{\rm SI}$ increases), a clear interpretation should certainly include considerations about the load, but also the variability of the flow rates and the traffic correlation at RTT scale. We leave it as future work.

7.2.3 Conclusion

In this section, we presented preliminary experimental results to assess the impact of the tail index of the flow-size distribution on QoS with the TCP protocol.

Based on experiments performed on our metrology platform, we showed that the vari-

ations of the QoS metrics with the tail index α_{SI} are more complex with TCP than they were with UDP. More precisely, we observed in our experiments that the mean packet delay and the mean throughput decrease when α_{SI} increases independently of the distribution used ((P1) or (P2)), but that the loss rate can decrease or remain constant depending on the distribution. We also observed that the flow-rate distribution is roughly insensitive to α_{SI} in most situations. Finally, we noted that, more than the tail index α_{SI}, the absolute position of the distribution for a large range of sizes might play an important role in the observed QoS.

However in our experiments, the results are likely to be mainly governed by long transient behaviors at the beginning of each flow, due to the absence of slow-start, and the results might be different in other experimental conditions. This experimental study shows the difficulty of reproducing real-world environments on controlled experimental platform. It also shows that complex TCP optimization parameters can deeply affect experimental results, which probably partially explains the existing gap between simplified TCP models and real-world observations.

Conclusion

In this chapter, we tackled the question of the impact of the flow-size distribution tail index on different Quality of Service metrics.

Based on traffic and queue simulations, we first showed that in the case of UDP traffic, α_{SI} systematically degrades the performance for large buffers whereas it has no impact on the loss rate and buffer occupancy for small buffers. We also observed that the critical buffer size marking the transition between these two regimes is different for these two metrics, and interpreted this result by their different statistical nature which induce different sensitivity to correlations at scales larger than the buffer size.

Then, based on experiments performed on our metrology platform, we showed that the evolution of the QoS metrics for TCP traffic is less clear. In particular, we observed that depending on the shape of the distribution, the impact of the tail parameter on the loss rate can be vary. Some of our results are in contradiction with previous results in the literature, and show that QoS can be very sensitive to complex TCP parameters. In particular, our results are likely to be conditioned to the absence of slow-start in our experiments, and need further work to be assessed in more generality and interpreted more deeply.

Impact of the traffic characteristics on QoS (in performance terms).

- In realistic conditions implying finite buffer sizes, what is the impact of heavy-tailed flow-size distributions on QoS (in performance terms)? More precisely, on which QoS metric?

With UDP, the loss rate and mean buffer occupancy are degraded with very heavy tails for large buffers, and are insensitive to α_{SI} for small buffers. The critical buffer size marking the transition between these two regimes is approximately the mean flow duration for the mean buffer occupancy, and is smaller for the loss rate.

With TCP, the results can depend more on the whole distribution than only on α_{SI}; and be very sensitive to complex TCP optimization parameters.

Chapter 8

Scaling laws in TCP traffic and throughput predictability

The results of this chapter have not yet been published.

TCP traffic at packet scale and TCP throughput predictability.

- Do TCP mechanisms generate muti-fractal scaling laws in the packet-level source traffic? What is their relation with TCP throughput prediction (beyond the mean)?

Introduction

In the last three chapters, we focused on the aggregate network traffic and its impact on QoS. From the viewpoint of a user who transmits a long flow, the particular traffic emitted by his own source is also very important to characterize. When the protocol is TCP, this corresponds to the characterization of the congestion window's evolution. We have seen in Section 3.2 of Chapter 3 that most of the works on this question have concentrated on the evaluation of the mean throughput of TCP traffic in various conditions (the almost-sure mean throughput given by the ergodic theorem (Theorem 2.1.7 of Chapter 2). To characterize more deeply the traffic variability, a few papers have attempted a multifractal description of the Hölder regularity, but none of them have conclusively linked this multifractal property to the TCP protocol behavior.

In this chapter, we use a different approach of mutifractal scaling laws. The scale invariant quantity that we consider is no longer the Hölder regularity, but the mean throughput in consecutive time windows of size n RTT. We show that this quantity exhibits a multifractal behavior genuinely generated by TCP control mechanisms, and we discuss the consequences of these new scaling laws in term of TCP throughput and fairness predictability. Our approach consists in looking at the probability that the mean throughput in time windows of size n RTT deviates from the almost-sure "global" mean. While large-deviations theory offers a precise estimate of this probability on a large number of independent realizations, we are interested here in the case of only one realization. We first provide in Section 8.1 an original large-deviation theorem that establishes the large-deviation spectrum observable on one realization of a general stochastic process. Section 8.1.4 summarizes the simplified result in the case Markov chains (that we use to model TCP traffic in the next section) and elaborates on its interpretation. In Section 8.2, we illustrate on a few example of real TCP

traces the applicability of our theoretical result, its contribution to the characterization of TCP throughput and its practical limits.

8.1 An almost-sure large-deviation result for one realization of a stationary ergodic process

In this section, we establish an almost-sure large-deviation theorem in a general mathematical context. The simplified results used for the application to TCP traffic (Section 8.2) are summarized at the end (Section 8.1.4).

8.1.1 Framework

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which all random variables are defined. We consider a discrete time process $X = (X_0 X_1 \cdots)$ taking values in a space $E^{\mathbb{N}}$ and we denote by μ its law, and by T the shift on $E^{\mathbb{N}}$:

$$T^j X = (X_j X_{j+1} \cdots).$$

In what follows, we will be interested in processes satisfying some mixing properties. There exists several mixing coefficients [Doukhan, 1994, Rio, 2000], which roughly describe how far from independent is a sequence of random variables. We introduce here only the strong mixing and uniform mixing coefficients that we use in the following. For two sub σ-algebras $\mathcal{U}$ and $\mathcal{V}$ of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the strong mixing coefficient is defined as:

$$\alpha(\mathcal{U}, \mathcal{V}) = \sup \left\{ \left| \mathbb{P}(U)\mathbb{P}(V) - \mathbb{P}(U \cap V) \right|, U \in \mathcal{U}, V \in \mathcal{V} \right\},$$

and the uniform mixing coefficient is defined as:

$$\phi(\mathcal{U}, \mathcal{V}) = \sup \left\{ \left| \mathbb{P}(V) - \frac{\mathbb{P}(U \cap V)}{\mathbb{P}(U)} \right|, U \in \mathcal{U}, \mathbb{P}(U) \neq 0, V \in \mathcal{V} \right\}.$$

The absolute value in these definitions ensures positivity of the mixing coefficients.

If we denote by $\chi(X_i)$ the σ-algebra generated by X_i (we do not use the classical notation $\sigma(X_i)$ to avoid confusion with variances in the following), the mixing coefficients of the sequence $(X_i)_{i \in \mathbb{N}}$ are defined by:

$$\alpha_m^{(X)} = \sup \left\{ \alpha \left(\chi(X_i), \chi(X_j) \right), (i,j) \in \mathbb{N}^2, |i-j| \geq m \right\}, \; m \geq 1, \tag{8.1}$$

$$\phi_m^{(X)} = \sup \left\{ \phi \left(\chi(X_i), \chi(X_j) \right), (i,j) \in \mathbb{N}^2, |i-j| \geq m \right\}, \; m \geq 1, \tag{8.2}$$

with $\alpha_0^{(X)} = \frac{1}{2}$ and $\phi_0^{(X)} = 1$. With these definitions, it is easily seen that the sequences of mixing coefficients are non-increasing.

The sequence $(X_i)_{i \in \mathbb{N}}$ is said ϕ-mixing or α-mixing if the corresponding mixing coefficient tends to 0 as m tends to infinity. Note that we always have:

$$\alpha_m^{(X)} \leq \frac{1}{2} \phi_m^{(X)}, \tag{8.3}$$

so that a ϕ-mixing sequence is also α-mixing.

We introduce a function $\Phi : E^{\mathbb{N}} \to \mathbb{R}^d, d \in \mathbb{N}^*$ and assume that Φ depends only on a finite number $l \geq 1$ of coordinates of X: $\Phi(X) = \Phi(X_0 \cdots X_{l-1})$. The function Φ can then be simply seen as a function from E^l to $\mathbb{R}^d$. For $q \in \mathbb{R}^d$ we define the following quantity:

$$P_n(q) = \frac{1}{n} \log \mathbb{E}_\mu \{ e^{\langle q, S_n \Phi \rangle} \}, \tag{8.4}$$

where $\langle \cdot, \cdot \rangle$ denotes the classical inner product on $\mathbb{R}^d$, and S_n denotes the Birkhoff sum: $S_n \Phi(X) = \sum_{j=0}^{n-1} \Phi(T^j X)$. It is well-known that for a large category of processes, this quantity converges when n goes to infinity to a function:

$$P(q) = \lim_{n \to \infty} P_n(q), \tag{8.5}$$

called pressure in the dynamical system context, partition function in the multifractal context, free energy in physics or logarithmic moment generating function in the large-deviation context. In this work, we adopt the denomination pressure (and use the notation P). It has been established that the differentiability of this function is a sufficient condition for an associated large-deviation principle to hold (this is the so called Gartner-Ellis theorem [Ellis, 1984]). This condition is known to hold for a number of processes (see [Dembo and Zeitouni, 1998]). However, it is also known to fail for certain ϕ-mixing processes with exponential rates [Bryc and Dembo, 1996, Baxter et al., 1991], and we will then take it as an assumption in the following. However, the pressure defined by equations (8.4) and (8.5) is based on an ensemble mean, which corresponds to the averaging on a large number of independent realizations of process X. In practical situations, we often have only one realization of the process, and the question of whether such a large-deviation principle holds with a pressure defined through an empirical mean on this realization is of primary interest. The answer to that question is given in Theorem 8.1.1, which constitutes our main result of this section.

8.1.2 An almost-sure large-deviation result for a uniformly mixing process

Statement

Theorem 8.1.1. *Assume that:*

(i, stationary) *X is a stationary ergodic process of law μ,*

(ii, mixing prop.) *the mixing coefficients of the sequence $(X_i)_{i \in \mathbb{N}^*}$ satisfy $\sum_{m=1}^{\infty} \phi_m^{(X)} < \infty$,*

(iii, LDP holds) *for all $q \in \mathbb{R}^d$, the pressure defined in equation (8.5) exists and is differentiable,*

(iv, Φ bounded) *$\forall q \in \mathbb{R}^d \; \exists M(q) < \infty$: $||\langle q, \Phi(X) \rangle||_\infty < M(q)$.*

Let K be a open convex subset of $\mathbb{R}^d$, and let $(k_n)_{n \in \mathbb{N}^*}$ be a sequence of integers such that:

$$\frac{\sup_{q \in K} \left\{ \left[1 \vee e^{n(M(q) - P_n(q))} \right]^2 \cdot \sum_{m=1}^{\infty} \phi_m^{(X)} \right\} \cdot \log(n)}{k_n} \xrightarrow[n \to \infty]{} 0 \tag{8.6}$$

Then, $\mu-$ a.s., for all $q \in K$:

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log \frac{\# \left\{ j \in \{0, \cdots, k_n - 1\} : \frac{S_n \Phi \circ T^{nj}(X)}{n} \in \mathcal{B}^o(\alpha, \varepsilon) \right\}}{k_n} = f(\alpha), \tag{8.7}$$

where $\alpha = \nabla P(q)$, $\mathcal{B}^o(\alpha, \varepsilon)$ is the open ball of radius ε centered on α and $f(\alpha) = -P^*(\alpha)$ is the opposite of the Legendre-Fenchel transform of P:

$$f(\alpha) = - \sup_{q \in \mathbb{R}^d} \{\langle q, \alpha \rangle - P(q)\} \tag{8.8}$$

Remark 8.1.2. The condition (ii) implies that the process X is mixing (recall that the mixing coefficients are positive).

Remark 8.1.3. For simplicity, we used the notation $\cdot \vee \cdot$ to denote the larger of two elements.

Proof

The key point of the proof is to show that the equivalent $L_n(q)$ of the pressure defined through an empirical mean on one realization converges toward the pressure $P(q)$, so that we can apply the Gartner-Ellis theorem with this new function. More precisely, we define:

$$L_n(q) = \frac{1}{n} \log \frac{1}{k_n} \sum_{j=0}^{k_n - 1} e^{\langle q, S_n \Phi \circ T^{nj}(X) \rangle}, \ \forall q \in \mathbb{R}^d, \tag{8.9}$$

which is the equivalent of the pressure of equation (8.4), where the ensemble mean has been replaced by an empirical mean over k_n windows of size n.

The proof operates in 3 steps:

1. We first show that for all $q \in \mathbb{R}^d$, $L_n(q)$ converges almost surely towards $P(q)$.
2. Using a result of convex analysis, we then show that almost surely, for all q in a bounded open set K, L_n converges toward P.
3. We finally apply Gartner-Ellis theorem to obtain the large-deviation spectrum of equation (8.7).

For all $q \in \mathbb{R}^d$, almost-sure convergence of $L_n(q)$

Lemma 8.1.4 (Convergence result)**.** *Let $q \in \mathbb{R}^d$. Suppose that assumptions (i)-(iv) of Theorem 8.1.1 are verified, and that k_n is a sequence of integers such that*

$$\frac{\left\{ \left[1 \vee e^{n(M(q) - P_n(q))} \right]^2 \cdot \sum_{m=1}^{\infty} \phi_m^{(X)} \right\} \cdot \log(n)}{k_n} \xrightarrow[n \to \infty]{} 0 \tag{8.10}$$

Then,

$$L_n(q) \xrightarrow[n \to \infty]{} P(q), \quad \mu - a.s. \tag{8.11}$$

Proof. We fix $q \in \mathbb{R}^d$. First note that by stationarity, we have

$$\forall n, \forall j \ : \ \mathbb{E}_\mu e^{\langle q, S_n \Phi \circ T^{nj}(X)\rangle} = e^{nP_n(q)}.$$

To simplify the calculations, we then define the centered random variable:

$$X_j^{(n)}(q) = e^{\langle q, S_n \Phi \circ T^{nj}(X)\rangle - nP_n(q)} - 1, \tag{8.12}$$

$$\mathbb{E}_\mu X_j^{(n)}(q) = 0.$$

Then,

$$L_n(q) = \frac{1}{n} \log \left[1 + \frac{\mathcal{S}_n(q)}{k_n}\right] + P_n(q), \tag{8.13}$$

where

$$\mathcal{S}_n(q) = \sum_{j=1}^{k_n} X_j^{(n)}(q). \tag{8.14}$$

To prove lemma 8.1.4, it suffices to prove that $\frac{\mathcal{S}_n(q)}{k_n}$ converges almost surely towards zero (note that it is not necessary).

We use the exponential inequality provided by Theorem 5 of [Doukhan, 1994, Section 1.4.2], applied to the sequence $(X_j^{(n)}(q))_{j\geq 0}$ for n fixed; which states that if $(X_j^{(n)}(q))_{j\geq 0}$ is a stationary centered process such that: (a) $(X_j^{(n)}(q))_{j\geq 0}$ is bounded by $M_n(q)$, and (b) the sequence is uniformly mixing with $\lim_{m\to\infty} m\phi_m^{(X^{(n)})} = 0$, then for all $x > 0$,

$$\mathbb{P}\left(|\mathcal{S}_n(q)| \geq x\sqrt{k_n}\right) \leq a \exp\left(-\frac{bx^2}{\sigma_n^2(q) + \frac{xM_n(q)\psi(\sigma_n^2(q))}{\sqrt{k_n}}}\right), \tag{8.15}$$

where $\sigma_n^2(q)$ is any real number satisfying:

$$\sigma_n^2(q) \geq \frac{\mathbb{V}ar\left[\sum_{j=1}^k X_j^{(n)}\right]}{k}, \ \forall k > 0,$$

and $\psi(t) = \inf\{m \in \mathbb{N}; m\phi_m^{(X^{(n)})} \leq t\}$. Moreover, the constants a and b are independent of $M_n(q)$ and $\sigma_n^2(q)$.

In our case, the bounded assumption (a) is directly deduced from the bounded hypothesis on Φ (assumption (iv) of Theorem 8.1.1), where the bound $M_n(q)$ is given by

$$\begin{aligned} ||X_j^{(n)}||_\infty &= ||e^{\langle q, S_n \Phi \circ T^{nj}(X)\rangle - nP_n(q)} - 1||_\infty \\ &\leq \left[1 \vee e^{n(M(q) - P_n(q))}\right] = M_n(q). \end{aligned} \tag{8.16}$$

The mixing condition (b) on the sequence $(X_j^{(n)}(q))_{j\geq 0}$ comes from the mixing condition (ii) of Theorem 8.1.1 and from the fact that $\Phi(X)$ depends only on a finite number l of coordinates of X. Indeed,

$$\phi_m^{(X^{(n)})} = \sup\left\{\phi\left(\chi(X_i^{(n)}), \chi(X_j^{(n)})\right), (i,j) \in \mathbb{N}^2, |i-j| \geq m\right\},$$

and $\chi(X_i^{(n)}) \subset \chi(X_{jn+1}, \cdots, X_{n(j+1)+l})$, so that

$$\phi_m^{(X^{(n)})} \leq \phi_{n(m-1)+1-l}^{(X)}, \tag{8.17}$$

where the index $n(m-1)+1-l$ is put to zero if negative (for small values of n and m). We then clearly have:

$$\sum_{m=1}^{\infty} \phi_m^{(X^{(n)})} < \infty, \tag{8.18}$$

which, using the assumption that the sequence of mixing coefficients is non-increasing (satisfied from the definition in equation (8.2)), implies that condition (b) is satisfied.

Finally, for the upper bound σ_n^2, we use the result of [Rio, 2000, page 14-15]:

$$\frac{\mathbb{V}ar\left[\sum_{j=1}^{k} X_j^{(n)}\right]}{k} \leq 2M_n(q)^2 \sum_{m\geq 0} \phi_m^{(X^{(n)})} = \sigma_n^2(q), \ \forall k > 0. \tag{8.19}$$

This result of [Rio, 2000] is proven in the case of a strongly mixing sequence of random variables bounded by $M_n(q)$ and concerns the sum of the strong mixing coefficient. We adapted it here to the case of uniform mixing (see equation (8.3)).

Let ε' be a positive real number. Equation (8.15), applied with $x = \varepsilon'\sqrt{k_n}$ brings:

$$\mathbb{P}\left(\frac{|\mathcal{S}_n(q)|}{k_n} > \varepsilon'\right) \leq a \exp\left(-\frac{b\varepsilon'^2 k_n}{\sigma_n^2(q) + \varepsilon' M_n(q)\psi\left(\sigma_n^2(q)\right)}\right). \tag{8.20}$$

Given the value of $\sigma_n^2(q) \sim M_n(q)^2$ that we have chosen in equation (8.19), condition (8.10) implies that there exists a function $f(n)$ which tends to infinity when n tends to infinity, such that:

$$\frac{k_n}{\sigma_n^2(q) + \varepsilon' M_n(q)\psi\left(\sigma_n^2(q)\right)} = \log(n) f(n). \tag{8.21}$$

Indeed, from equations (8.19) and (8.16), the term in braces in equation (8.10) is $\sigma_n^2(q)$ (where the constant factor 2 was removed). Moreover, since $M_n(q)$ is exponentially increasing with n (see equation (8.16) and note that $P_n(q) \leq M(q)$), $\sigma_n^2(q)$ is non-decreasing when n increases and ψ is then easily bounded from above. Indeed, since $\lim_{m\to\infty} m\phi_m^{(X^{(n)})} = 0$, we can find m_0 such that $m_0\phi_{m_0}^{(X^{(n)})} \leq \sigma_1^2(q)$ and thus $\psi(\sigma_1^2(q)) \leq m_0$. Then, for any integer n, $\sigma_n^2(q) \geq \sigma_1^2(q)$, so that $m_0\phi_{m_0}^{(X^{(n)})} \leq \sigma_n^2(q)$ and $\psi(\sigma_n^2(q)) \leq m_0$. Finally, since $\sigma_n^2(q) \sim M_n(q)^2$, in the denominator of the fraction of equation (8.21), we have $\varepsilon' M_n(q)\psi\left(\sigma_n^2(q)\right) = o(\sigma_n^2(q))$, so that condition (8.10), where only the term $\sigma_n^2(q)$ is "dominated", ensures validity of equation (8.21).

With a constant C depending on ε' and b, equations (8.20) and (8.21) imply that:

$$\mathbb{P}\left(\frac{|\mathcal{S}_n(q)|}{k_n} > \varepsilon'\right) \leq an^{-Cf(n)}.$$

Since, $f(n)$ goes to infinity, there exists an integer n_0 such that $Cf(n) > 1, \forall n > n_0$, which ensures the convergence of the series $\sum_{n\geq 0} an^{-Cf(n)}$. By Borel-Cantelli's lemma, we then have:

$$\frac{\mathcal{S}_n(q)}{k_n} \xrightarrow[n\to\infty]{} 0, \quad \mu - \text{a.s.}, \tag{8.22}$$

which, given equation (8.13), completes the proof of lemma 8.1.4. □

Remark 8.1.5. We see from equation (8.20) that to ensure this convergence result, k_n must roughly "dominate" (in the sense that a relation as equation (8.21) can be written) two term: $\sigma_n^2(q)$ (the variance term) and $\varepsilon' M_n(q)$ (the maximum term). With the upper bound of equation (8.19) that we have taken as a function of $M_n(q)^2$, domination of the term $\sigma_n^2(q)$ automatically implies domination of the term with $M_n(q)$. However, this is a rough upper bound of $\frac{\mathbb{V}ar\left[\sum_{j=1}^k X_j^{(n)}\right]}{k}$. In most cases, the critical term to dominate in equation (8.20) is the variance term, and we shall see in next section a finer bound for this term, based on the pressure. The precise conditions on the distribution of $X^{(n)}$ for this to hold are still under investigation, but intuitively, we see that since $\mathbb{E}X_j^{(n)} = 0$, it should not be difficult to find a constant C such that $M_n(q) \leq C\sigma_n^2(q)$ for any upper bound $\sigma_n^2(q)$ of $\frac{\mathbb{V}ar\mathcal{S}_n(q)}{k_n}$, so that domination of the term $\sigma_n^2(q)$ automatically implies domination of the term $\varepsilon' M_n$. When choosing a tighter bound $\sigma_n^2(q)$ for $\frac{\mathbb{V}ar\left[\sum_{j=1}^k X_j^{(n)}\right]}{k}$, one must ensure that the value of ψ does not introduce problems. However, the variance of the terms $X_j^{(n)}(q)$ constituting the sum is exponentially increasing as $e^{nR_n(q)}$, with $R_n(q) = P_n(2q) - 2P_n(q) \geq 0$ (see next section); so that making σ_n^2 a non-decreasing function of n (and using arguments mentioned above) should not be a big constraint.

Convexity argument: Almost surely, for all $q \in K$ convergence of $L_n(q)$.

Let K be an open convex subset of $\mathbb{R}^d$ and K_d a countable dense subset of K. For all $q \in K_d$, under assumption (8.6), $L_n(q)$ converges toward $P(q)$ outside a negligible subset N_q of Ω. The countable union $N = \cup_{q \in K_d} N_q$ is still negligible, so that almost surely, $L_n(q)$ converges toward $P(q)$ for all $q \in K_d$. Since the functions L_n are convex, Theorem 10.8 of [Rockafellar, 1970] ensures that outside the negligible set N (*i.e.* μ-a.s.), $L_n(q)$ converges towards $P(q)$ for all q in K.

Application of the Gartner-Ellis Theorem

The last part of the proof follows the lines of a classical proof in the context of Hölder multifractality (see [Riedi, 2003]). We denote by $\mathbb{E}_k$ the expectation with respect to a uniform choice of an interval of size n among the k_n intervals. We have, μ – a.s., for all $q \in K$;

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_k\{e^{\langle q, S_n\Phi(X)\rangle}\} = P(q), \tag{8.23}$$

where function P is differentiable by assumption (iii) of Theorem 8.1.1.

By the Gartner-Ellis Theorem [Ellis, 1984], we then have μ – a.s., $\forall q \in \mathbb{R}^d$ and for $\alpha = \nabla P(q)$,

$$\limsup_{n\to\infty} \frac{1}{n} \log Q_n\{\mathcal{B}^c(\alpha, \varepsilon)\} \leq \sup_{\beta \in \mathcal{B}^c(\alpha,\varepsilon)} f(\beta), \tag{8.24}$$

and

$$\liminf_{n\to\infty} \frac{1}{n} \log Q_n\{\mathcal{B}^o(\alpha, \varepsilon)\} \geq \sup_{\beta \in \mathcal{B}^o(\alpha,\varepsilon)} f(\beta), \tag{8.25}$$

where Q_n is the distribution of $\frac{S_n\Phi(X)}{n}$ with respect to the uniform measure between 1 and k_n, and $\mathcal{B}^c(\alpha, \varepsilon)$ is the closed ball of radius ε centered on α.

Since for any n, α, ε, we have $Q_n\{\mathcal{B}^c(\alpha,\varepsilon)\} \geq Q_n\{\mathcal{B}^o(\alpha,\varepsilon)\}$, the upper bound of equation (8.24) can be written:

$$\limsup_{n\to\infty} \frac{1}{n} \log Q_n\{\mathcal{B}^o(\alpha,\varepsilon)\} \leq \limsup_{n\to\infty} \frac{1}{n} \log Q_n\{\mathcal{B}^c(\alpha,\varepsilon)\} \leq \sup_{\beta\in\mathcal{B}^c(\alpha,\varepsilon)} f(\beta),$$

which gives:

$$\sup_{\beta\in\mathcal{B}^o(\alpha,\varepsilon)} f(\beta) \leq \liminf_{n\to\infty} \frac{1}{n} \log Q_n\{\mathcal{B}^o(\alpha,\varepsilon)\} \leq \limsup_{n\to\infty} \frac{1}{n} \log Q_n\{\mathcal{B}^o(\alpha,\varepsilon)\} \leq \sup_{\beta\in\mathcal{B}^c(\alpha,\varepsilon)} f(\beta).$$

Now, by continuity of f, the two extreme terms of the inequality are equal, so that we have:

$$\lim_{n\to\infty} \frac{1}{n} \log Q_n\{\mathcal{B}^o(\alpha,\varepsilon)\} = \sup_{\beta\in\mathcal{B}^o(\alpha,\varepsilon)} f(\beta) \tag{8.26}$$

By continuity of f again, we have:

$$\lim_{\varepsilon\to 0} \sup_{\beta\in\mathcal{B}^o(\alpha,\varepsilon)} f(\beta) = f(\alpha),$$

so that taking the limit $\varepsilon \to 0$ in equation (8.26) gives the result of equation (8.7) and completes the proof of Theorem 8.1.1.

8.1.3 The particular case of Markov chains

The assumptions of Theorem 8.1.1 are fairly general and a large class of processes enter this framework. Independent processes satisfy these assumptions, but also ergodic Gibbs measures (ergodic Gibbs measures are ϕ-mixing with geometrically decreasing mixing coefficients and the pressure is analytic, see [Bowen, 1975, Ruelle, 1984, Zinsmeister, 2000]).

In this section, we consider the special case of Markov chains, particularly useful to model TCP traffic. For Markov chains, the ϕ-mixing property is equivalent to the irreducible aperiodic property; and the mixing sequence then decreases exponentially [Bradley, 2005], so that the mixing assumption (ii) of Theorem 8.1.1 is satisfied for an irreducible aperiodic Markov chain. We first give another condition on the sequence $(k_n)_{n\in\mathbb{N}}$ for the result of Theorem 8.1.1 to hold. This condition relies on the computation of the pressure and we then give a simple procedure to compute this pressure, based on the computation of the spectral radius of a matrix derived from the transition matrix of the Markov chain.

In all the rest of this chapter, the process X is an irreducible aperiodic Markov chain, whose transition matrix is denoted by Q. The state space is assumed finite of cardinal c and we denote it $E = \{e_1, \cdots, e_c\}$. Note that the states e_i can be vectors.

A practical condition on the sequence $(k_n)_{n\in\mathbb{N}}$

As we have noted in Remark 8.1.5, the critical condition on the sequence $(k_n)_{n\in\mathbb{N}}$ is to dominate the term $\sigma_n^2(q)$ (in equation (8.20)), which is an upper bound of $\frac{\mathbb{V}ar\mathcal{S}_n(q)}{k_n}$. We give here another upper bound than the one used in previous section (equation (8.19)), and give the associated condition on k_n.

Let us first introduce the following quantities:

Definition 8.1.6. For $q \in \mathbb{R}^d$, we define:

$$R_n(q) = P_n(2q) - 2P_n(q), \tag{8.27}$$

$$R(q) = \lim_{n\to\infty} R_n(q) = P(2q) - 2P(q). \tag{8.28}$$

Remark 8.1.7. Convergence of R_n follows from convergence of P_n.

Remark 8.1.8. Since the function P is convex and $P(0) = 0$, $R(q)$ is non negative for all $q \in \mathbb{R}^d$.

The following proposition gives a practical condition on the sequence $(k_n)_{n\in\mathbb{N}}$ for the result of Theorem 8.1.1 to hold:

Proposition 8.1.9. *If*

$$k_n = e^{(1+\delta)nR(q)}, \text{ for some } \delta > 0, \tag{8.29}$$

then the result of Theorem 8.1.1 holds.

Proof. We use the following upper bound of $\frac{\mathbb{V}ar\mathcal{S}_n(q)}{k_n}$ (recall that the process X is stationary):

$$\sigma_n^2(q) = \mathbb{V}ar_\mu\{X_0^{(n)}\} + \sum_{j=1}^{\infty} \left|\mathbb{C}ov_\mu\{X_0^{(n)}, X_j^{(n)}\}\right|. \tag{8.30}$$

First note that

$$\mathbb{V}ar_\mu\{X_0^{(n)}\} = e^{nR_n(q)} - 1.$$

For n sufficiently large (larger than the number of coordinates l on which depends Φ), since X is a Markov chain, only the first term of the sum in equation (8.30) is non-zero, so that

$$\sigma_n^2(q) = e^{nR_n(q)} + O\left(e^{nR_n(q)}\right) \leq Ce^{nR_n(q)},$$

for some constant C and n sufficiently large.

If $k_n = e^{(1+\delta)nR(q)}$, for some $\delta > 0$, then

$$\frac{\sigma_n^2(q)\cdot\log(n)}{k_n} \leq Ce^{n(R_n(q)-R(q)-\delta R(q))}\cdot\log(n).$$

Since $R_n(q) \xrightarrow[n\to\infty]{} R(q)$, there exists a integer N_0 such that for $n > N_0$, $|R_n(q) - R(q)| < \delta R(q)$, and then $(R_n(q) - R(q) - \delta R(q)) > 0$. Finally, we get that

$$\frac{\sigma_n^2(q)\cdot\log(n)}{k_n} \xrightarrow[n\to\infty]{} 0, \tag{8.31}$$

which ensures convergence of the right-hand-side sum of equation (8.20). We conclude in the same way as for the proof of Theorem 8.1.1. □

The condition on the sequence $(k_n)_{n\in\mathbb{N}}$ of Proposition 8.1.9 will be used in the rest of this chapter. It relies on quantity $R(q)$, which is directly derived from the pressure. We now give a practical procedure to compute this pressure in the case of irreducible aperiodic Markov chains.

Computation of the pressure for an irreducible aperiodic Markov chain.

To give a practical procedure to compute the pressure (Proposition 8.1.10), we restrict ourselves to the case where the function Φ depends only on the first coordinate X_1 of X ($l = 1$); Φ is then a function from E to $\mathbb{R}^d$. The case $l \geq 1$ can be handled in a similar manner, but yields a more complex result that we will not use in the following.

Proposition 8.1.10. *If X is an irreducible aperiodic Markov chain on a finite state space $E = \{e_1, \cdots, e_c\}$ of transition matrix Q, and $\Phi : E \to \mathbb{R}^d$; then $\forall q \in \mathbb{R}^d$,*

$$P(q) = \log\left(\rho\left(D(q)Q^t\right)\right), \tag{8.32}$$

where $D(q)$ is a diagonal matrix with the elements $\{\exp(\langle q, \Phi(e_1)\rangle), \exp(\langle q, \Phi(e_2)\rangle), \cdots, \exp(\langle q, \Phi(e_c)\rangle)\}$, Q^t denotes the transpose of Q and $\rho(\cdot)$ denotes the spectral radius.

Proof. We use the same method as Section 7.1 of [Fan and Feng, 2000], extended to the case of Markov measures. For simplicity, we introduce the following intuitive notations: $p_{e_i e_j} = Q_{ij}$ is the probability of a transition from state e_i to state e_j, and p_{e_i} is the steady-state probability of state e_i (note that the steady-state probability exists and is unique because we consider an irreducible Markov chain).

For $x \in E^{\mathbb{N}}$ and $q \in \mathbb{R}^d$, let

$$g(x) = \exp\langle q, \Phi(x)\rangle. \tag{8.33}$$

Then

$$P(q) = \lim_{n\to\infty} \frac{1}{n} \log\left[\int_{E^{\mathbb{N}}} \prod_{j=0}^{n-1} g(T^j x)\mu(x)\right]. \tag{8.34}$$

We write:

$$\int_{E^{\mathbb{N}}} \prod_{j=0}^{n-1} g(T^j x)\mu(x) = \sum_{x_1\cdots x_n \in E^n} g(x_1)p_{x_1} \cdot g(x_2)p_{x_1x_2} \cdots g(x_n)p_{x_{n-1}x_n}.$$

Let

$$\begin{aligned}
G_n(1) &= \sum_{x_1\cdots x_{n-1}\in E^{n-1}} g(x_1)p_{x_1} \cdot g(x_2)p_{x_1x_2} \cdots g(x_{n-1})p_{x_{n-2}x_{n-1}} \cdot g(1)p_{x_{n-1}e_1} \\
&\vdots \\
G_n(c) &= \sum_{x_1\cdots x_{n-1}\in E^{n-1}} g(x_1)p_{x_1} \cdot g(x_2)p_{x_1x_2} \cdots g(x_{n-1})p_{x_{n-2}x_{n-1}} \cdot g(c)p_{x_{n-1}e_c}.
\end{aligned}$$

Then we have the following recursive relation:

$$\begin{aligned}
G_{n+1}(1) &= G_n(1)g(e_1)p_{e_1e_1} + \cdots + G_n(c)g(e_1)p_{e_ce_1} \\
&\vdots \\
G_{n+1}(c) &= G_n(1)g(e_c)p_{e_1e_c} + \cdots + G_n(c)g(e_c)p_{e_ce_c},
\end{aligned}$$

which can be written:

$$\begin{pmatrix} G_{n+1}(1) \\ \vdots \\ G_{n+1}(c) \end{pmatrix} = A \begin{pmatrix} G_n(1) \\ \vdots \\ G_n(c) \end{pmatrix},$$

where

$$A = A(q) = \begin{pmatrix} g(e_1)p_{e_1e_1} & \cdots & g(e_1)p_{e_ce_1} \\ \vdots & & \vdots \\ g(e_c)p_{e_1e_c} & \cdots & g(e_c)p_{e_ce_c} \end{pmatrix} = \begin{pmatrix} e^{\langle q,\Phi(e_1)\rangle}p_{e_1e_1} & \cdots & e^{\langle q,\Phi(e_1)\rangle}p_{e_ce_1} \\ \vdots & & \vdots \\ e^{\langle q,\Phi(e_c)\rangle}p_{e_1e_c} & \cdots & e^{\langle q,\Phi(e_c)\rangle}p_{e_ce_c} \end{pmatrix}.$$

Thus, we have

$$\begin{pmatrix} G_n(1) \\ \vdots \\ G_n(c) \end{pmatrix} = A^{n-1} \begin{pmatrix} G_1(1) \\ \vdots \\ G_1(c) \end{pmatrix},$$

and since

$$\begin{pmatrix} G_1(1) \\ \vdots \\ G_1(c) \end{pmatrix} = A \begin{pmatrix} p_1 \\ \vdots \\ p_c \end{pmatrix},$$

we can finally write:

$$\begin{pmatrix} G_n(1) \\ \vdots \\ G_n(c) \end{pmatrix} = A^n \begin{pmatrix} p_{e_1} \\ \vdots \\ p_{e_c} \end{pmatrix}.$$

So,

$$\int_{E^{\mathbb{N}}} \prod_{j=0}^{n-1} g(T^j x)\mu(x) = \left\| A^n \begin{pmatrix} p_{e_1} \\ \vdots \\ p_{e_c} \end{pmatrix} \right\|_1,$$

where $||\cdot||_1$ denotes the norm of a vector defined by the sum of all the absolute values of its entries (note that, here, we only obtain the sum of all the entries but this is equivalent because all the entries are positive).

Let

$$A_j^{(n)} = \sum_{i=1}^{c} (A^n)_{ij}$$

be the sum of the elements of the jth column of the matrix A^n. Then,

$$\left\| A^n \begin{pmatrix} p_{e_1} \\ \vdots \\ p_{e_c} \end{pmatrix} \right\|_1 = \sum_{j=1}^{c} A_j^{(n)} p_{e_j}.$$

Let $p_{\min} = \min_{j=1,\cdots,c} p_{e_j}$ and $p_{\max} = \max_{j=1,\cdots,c} p_{e_j}$ By irreducibility of the Markov chain, we have $0 < p_{\min} < 1$ and $0 < p_{\max} < 1$.

We now write:

$$\sum_{j=1}^{c} A_j^{(n)} p_{\min} \leq \left\| A^n \begin{pmatrix} p_{e_1} \\ \vdots \\ p_{e_c} \end{pmatrix} \right\|_1 \leq \sum_{j=1}^{c} A_j^{(n)} p_{\max},$$

and then:

$$p_{\min}||A^n||_1 \leq \left\| A^n \begin{pmatrix} p_{e_1} \\ \vdots \\ p_{e_c} \end{pmatrix} \right\|_1 \leq p_{\max}||A^n||_1$$

So,

$$\log\left[||A^n||_1^{\frac{1}{n}}\right] + \frac{1}{n}\log(p_{\min}) \leq \frac{1}{n}\log\left[\left\|A^n\begin{pmatrix} p_{e_1} \\ \vdots \\ p_{e_c}\end{pmatrix}\right\|_1\right] \leq \log\left[||A^n||_1^{\frac{1}{n}}\right] + \frac{1}{n}\log(p_{\max}),$$

and finally, by Gelfand's formula,

$$P(q) = \log(\rho(A(q))), \tag{8.35}$$

where $\rho(A(q))$ denotes the spectral radius of $A(q)$. Noting that $A(q) = D(q)Q^t$, where $D(q)$ is a diagonal matrix with the elements $\{\exp(\langle q, \Phi(e_1)\rangle), \exp(\langle q, \Phi(e_2)\rangle), \cdots, \exp(\langle q, \Phi(e_c)\rangle)\}$ completes the proof. □

8.1.4 Simplified results used in the following and interpretation

We summarize here the main result, in a simplified form used in the following for the application to TCP traffic analysis, and elaborate on its interpretation. For more generality, Theorem 8.1.1 has been proven for a function Φ taking values in $\mathbb{R}^d$, and depending on a finite number l of coordinates of the process X. In the following, we consider only the case $d = 1$ and $l = 1$ (*i.e.* Φ is a function from E to $\mathbb{R}$). The values of q and α are then in $\mathbb{R}$, and the open ball of center α and of radius ε then corresponds to the interval $(\alpha - \varepsilon, \alpha + \varepsilon)$. Moreover, we consider the case of an irreducible aperiodic Markov chain on a finite state space. Then, we use the following result, which is a straightforward corollary of Theorem 8.1.1, Proposition 8.1.9 and Proposition 8.1.10:

Corollary 8.1.11. *Let X be an irreducible aperiodic Markov chain on finite-state space $E = \{e_1, \cdots, e_c\}, c \geq 1$, and let Φ be a function $E \to \mathbb{R}$.*

Define the quantity

$$\gamma_j^{(n)} = \frac{\sum_{i=jn+1}^{(j+1)n} \Phi(X_i)}{n}, \tag{8.36}$$

corresponding to the sample mean of $\Phi(X)$ on the jth interval of size n.

Then for all $q \in \mathbb{R}$, provided that the number $k_n(q)$ of intervals of size n is large enough for each n in the sense:

$$k_n(q) = e^{(1+\delta)nR(q)}, \text{ for some } \delta > 0, \tag{8.37}$$

the following large-deviation result holds on almost-every realization of the process X:

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log\frac{\#\left\{j \in \{0, \cdots, k_n - 1\} : \gamma_j^{(n)} \in (\alpha - \varepsilon, \alpha + \varepsilon)\right\}}{k_n} = f(\alpha), \tag{8.38}$$

where $\alpha = P'(q)$, $f(\alpha)$ is the opposite of the Legendre-Fenchel transform of P:

$$f(\alpha) = -\sup_{q\in\mathbb{R}}\{\alpha q - P(q)\},$$

and $P(q)$ can be computed as:

$$P(q) = \log\left(\rho\left(A(q)\right)\right), \tag{8.39}$$

where $A(q)$ is a $c \times c$ matrix of elements $(A(q))_{i,j} = e^{q\cdot\Phi(e_i)}Q_{j,i}$, and $\rho(\cdot)$ denotes the spectral radius, i.e. *the largest absolute value of the eigenvalues.*

Corollary 8.1.11 establishes that a large-deviation principle holds on one realization of the process X: roughly, for a given value α, the proportion of values of the sample mean of $\Phi(X)$ ($\gamma_j^{(n)}$) around α decreases as $\exp(nf(\alpha))$ (note that $f(\alpha)$ is always non-positive with the definition of equation (8.38)). We call the function $f(\alpha)$ the *large-deviation spectrum* (as for the pressure, it also admits several names, depending on the context). It does not depend on n, which reflects the scale invariance of the repartition of $\gamma_j^{(n)}$ and is thus characteristic of the presence of a *scaling law*.

The possible values of α are those for which $\alpha = P'(q)$ for some $q \in \mathbb{R}$ (for other values of α, $f(\alpha)$ is usually set to $-\infty$ and the proportion of interval on which $\gamma_j^{(n)}$ then takes this value is zero: such values of the sample mean are not observable). In many cases (see next section), the pressure is strictly convex (*i.e.* non-linear), so that there is an ensemble of possible values of α for which $f(\alpha) > -\infty$. The scaling law is then *multifractal* (the spectrum is said *non-degenerate*, *i.e.* not reduced to a single point), and there is a one-to-one correspondence between values of α and q given by $\alpha(q) = P'(q)$. In the case $q = 0$, the corresponding value of α is the almost-sure mean value given by the ergodic theorem (Theorem 2.1.7 of Chapter 2); and the large-deviation spectrum takes the value $f(\alpha) = 0$. Positive values of q correspond to values of α larger than the almost-sure mean and negative values of q correspond to values of α smaller than the almost-sure mean (according to the convexity of P, which implies that its derivative is increasing).

In practice, to verify the scale invariance of the statistical repartition of the quantities $\gamma_j^{(n)}$, one computes an estimation of the large-deviation spectrum at different scales n (here, we use an original estimation method proposed in [Barral and Gonçalves, 2009]). Then, a superimposition of the empirical spectra at the different scales indicates that this repartition does not depend on the scale (it validates the convergence in equation (8.38)).

For Markov chains, a classical large-deviation principle is known to hold (see *e.g.* [Dembo and Zeitouni, 1998]). It establishes the exponential decrease of the (ensemble) probability that, over an interval of size n, the sample mean $\gamma_0^{(n)}$ is around a value α; that is the proportion of such intervals, computed on an infinite number of independent realizations of the process. In this context, the originality of Corollary 8.1.11 is that it establishes that a large-deviation principle holds true on almost-every realization of the process taken separately, and with the same large-deviation spectrum. Thus, the multifractal structure can be observed on almost-every realization of the process X. Beyond the fact that it is a nice mathematical observation (the ensemble multifractal structure is carried out on almost-every single realization), the result of Corollary 8.1.11 is the necessary theoretical step to ensure consistency with application to TCP traffic, where the evolution of the TCP congestion window of a flow is described by one realization of a Markov chain.

There is however a "price to pay" to observe large deviations on one realization: the number k_n of intervals of size n has to grow exponentially fast with the scale n (condition (8.37)). For a value of q in $\mathbb{R}$, the part of the spectrum corresponding to the value $\alpha = P'(q)$ can be observed if $k_n(q)$ grows exponentially fast with a rate $R(q) = P(2q) - 2P(q)$. This rate increases with $|q|$, that is as the corresponding value of α deviates from the almost-sure mean value: "larger deviations (from the almost-sure mean) are more difficult to observe".

Practically, for a finite-size realization of size N, and for a finite scale n, one has $\lfloor N/n \rfloor$ intervals of size n. Then, we define the minimal and maximal values of q for which this

number of intervals is "sufficient" in the sense of condition (8.37):

$$\left\{ \begin{array}{rcl} q_{\min}(n) & = & \min\{q \in \mathbb{R}, \lfloor N/n \rfloor \geq k_n(q)\}, \\ q_{\max}(n) & = & \max\{q \in \mathbb{R}, \lfloor N/n \rfloor \geq k_n(q)\}, \end{array} \right. \tag{8.40}$$

and the corresponding values of α:

$$\left\{ \begin{array}{rcl} \alpha_{\min}(n) & = & P'(q_{\min}(n)), \\ \alpha_{\max}(n) & = & P'(q_{\max}(n)). \end{array} \right. \tag{8.41}$$

Roughly, within these bounds, the spectrum defined in equation (8.38) is observable at scale n from one realization of size N. In other terms, they determine the interval $[\alpha_{\min}, \alpha_{\max}]$ of observable sample means of $\Phi(X)$. For a given size N, this interval shrinks around the almost-sure mean as the scale n increases; while for a given scale n, it gets wider as the size of the realization N increases. Numerical examples of these bounds will be given in next section in the case of a Markov chain modeling TCP traffic.

8.1.5 Conclusion

In this section, we provided an original large-deviation theorem valid on almost-every realization of a mixing process. In the particular case of a Markov chain X, this theorem shows that the sample mean of $\Phi(X)$ in consecutive intervals of size n follows a multifractal scaling law, where Φ is an arbitrary function providing extra versatility. The originality of our result, in the context of Markov chain, lies in that the corresponding large-deviation spectrum is observable on one realization, provided that the number of intervals of size n grows sufficiently fast with the scale n. This condition also provides bounds to predict the part of the spectrum observable on a finite size realization of process X, and we finally provided a practical procedure to compute the theoretical large-deviation spectrum and its bounds for a given transition matrix.

Although further theoretical work would be required to give a more general statement of our theorem and find the less restrictive conditions of its validity, its application to the simple case of Markov chains already suffices to demonstrate the existence of new scaling laws in TCP traffic, closely related to performance issues. We now turn to this aspect.

8.2 Application to TCP traffic

In this section, we use the theorem proven in previous section, applied to a Markov chain modeling TCP throughput, to demonstrate that TCP traffic exhibits multifractal scaling laws of large-deviation spectrum given by equation (8.38). We experimentally illustrate our results on a few examples of long-lived TCP flows, and discuss their consequences in term of TCP performance prediction.

8.2.1 A Markovian model of TCP throughput

Markov chains have been widely used to model the evolution of TCP congestion window's evolution (see Section 3.3 of Chapter 3). While many of these models aimed at including different TCP mechanisms like slow start or timeout, our goal here is to provide a model analytically supporting the observation of new scaling laws originating from the AIMD mechanism, and we do not include these other mechanisms.

We use a Markov chain $(X_i)_{i\geq 0}$, where X_i represents the congestion window value (in packets) at time i (that is after the ith RTT). This Markov chain is, as we will see, irreducible and aperiodic. Moreover, since there is always a maximal congestion windows (imposed either by the sender or the receiver), its state space E is finite.

To specify more precisely our model, we now explicit the structure of the transition matrix Q in the case of a TCP Reno connection. Our model has two separate components: the specifications of the AIMD mechanism, which impose the possible transitions, and the experimental conditions, which dictate the probability of these transitions.

For a single TCP connection, the state space of the Markov chain is simply $E = \{1, \cdots, c\}$. The simple AIMD mechanism considered here works as follows. If no loss is detected, the congestion window is increased by one, and if a loss is detected, it is divided by two. Instead of assuming Bernoulli losses, we assume that the loss probability is a function $p(\cdot)$ of the current congestion window. The non-zero coefficients of the transition matrix read, $\forall i \in \{1, \cdots, c\}$:

$$\left\{ \begin{array}{rcl} Q_{i,\min(i+1,c)} & = & 1 - p(i), \\ Q_{i,\max(\lfloor i/2 \rfloor,1)} & = & p(i). \end{array} \right. \tag{8.42}$$

The loss function $p(\cdot)$ represents the network conditions. Setting the loss probability to a constant function $p(\cdot)$ of the congestion window only is a simplifying assumption for it supposes that the loss probabilities are stationary and do not depend on the past. However, any function $p(\cdot)$ can be used, if it is consistent with the irreducibility of the Markov chain (which in realistic cases and with an appropriate state space is a very weak constraint). It can include many cases, such as Bernoulli losses (with a function of the form $p(i) = 1-(1-p_{\text{pkt}})^i$), or congestion losses when the TCP connection shares the bottleneck with some cross-traffic (this case will be developed in the following); and thus gives a large versatility to our model. However, its most important strength probably lies in its ability to adapt to the case of multiple competing connections. Although an arbitrary number of connections could theoretically be taken into account (at the cost of a numerical burden due to a large state space), we describe here only the case of two flows.

To handle that case, we consider now vectorial states:

$$X_i = \left(X_i^{(1)}, X_i^{(2)} \right), \tag{8.43}$$

where $X_i^{(1)}$ (respectively $X_i^{(2)}$) represents the congestion window of connection 1 (respectively 2) at time i. The state space is then $E = \{1, \cdots, c\}^2$. For simplicity, we assume that for each flow, the loss probability depends only on the sum $X_i^{(\text{sum})} = X_i^{(1)} + X_i^{(2)}$ of the current congestion windows, and we continue to denote by $p(\cdot)$ this probability. Given the value of the sum $X_i^{(\text{sum})}$, the loss probability of each flow is then independent of its individual congestion window value. Consequences of these assumptions will be discussed in Section 8.2.2. Let us mention though that this assumption is not a pre-requisite of the model. Indeed, any transition probability could be set when filling the transition matrix Q. However, this assumption allows us to include essential parameters of the system in a simple and efficient manner.

To complete the specification of the model, we need to precise the joint loss probability of the two flows. Let us denote by $l_i^{(f)}, f = 1, 2$ the boolean variable equal to 1 if flow f experiences a loss at time i. To include the effect of synchronization that tends to

increase the probability that both flows experience a loss at the same time, we introduce the *synchronization rate* $\text{syn} \in [0,1]$, and assume that the joint probabilities are as follows:

$$\left\{ \begin{array}{lcl} P(l_i^{(1)}=0, l_i^{(2)}=0 | X_i^{(\text{sum})}=s) & = & (1-p(s))^2 + \text{syn} \cdot p(s)(1-p(s)), \\ P(l_i^{(1)}=1, l_i^{(2)}=0 | X_i^{(\text{sum})}=s) & = & p(s)(1-p(s)) - \text{syn} \cdot p(s)(1-p(s)), \\ P(l_i^{(1)}=1, l_i^{(2)}=1 | X_i^{(\text{sum})}=s) & = & p(s)^2 + \text{syn} \cdot p(s)(1-p(s)). \end{array} \right. \quad (8.44)$$

The syn parameter is intended to describe the synchronization of the two sources, *i.e.* the tendency that they behave similarly (experience simultaneous losses or not). Special values of the synchronization rate defined this way lead to particular cases of interest: if $\text{syn}=0$, we recover the independent case where the probabilities of equation (8.44) are $(1-p(s))^2$, $p(s)(1-p(s))$ and $p(s)^2$; whereas $\text{syn}=1$ corresponds to the fully synchronized case where losses always affect both flows together (the probabilities of equation (8.44) are then $1-p(s)$, 0 and $p(s)$). Other definitions of the synchronization rate have been proposed (notably in [Baccelli and Hong, 2002]) and we do not claim our definition to be "universally better" than the other proposed definitions. However, we believe that its intuitive soundness and easy interpretation justify its introduction in our work.

With this complete specification of the loss probabilities, the transition matrix Q is easily filled according to the possible transitions dictated by the AIMD mechanism governing each flow. In the case of two flows, we will be interested in the behavior of the congestion window of one flow, but also in the sum of both congestion windows (representing the cumulated throughput of the two flows), and in the difference between the two congestion windows, because this last quantity is representative of the fairness. All of these cases can easily be handled by an appropriate choice of function Φ, as shown in Table 8.1. For all of these quantities, Corollary 8.1.11 ensures that the mean value in consecutive time windows of size n RTT (the exponent $\gamma^{(n)}$ of equation (8.36)) is associated with a multifractal scaling law, for which it gives the analytical formula of the large-deviation spectrum, easily computable with the help of Proposition 8.1.10 (with the transition matrix Q given by our model). As we have seen, this large-deviation spectrum reflects the distribution of the quantity $\gamma^{(n)}$, and thus allows predicting bounds on the achieved performance, in probabilistic terms. Such bounds give, *e.g.*, the probability that the mean throughput of a connection is around a certain value over a time window of n RTT; or the probability that, over a time window of n RTT, the throughput of one connection is twice the throughput of the other connection. The shape of the large-deviation spectrum is then of primary importance, for it allows to perform a *multi-scale analysis* of TCP performance parameters.

8.2.2 Experimental validation and applications

We now use the experimental facility described in Section 4.2 of Chapter 4 to experimentally demonstrate the ability of the proposed Markov model to accurately reproduce the newly defined large-deviation spectrum observed on real TCP traffic. We treat the case of one, then two TCP flows sharing a single bottleneck with either constant or random UDP cros-traffic. Empirical spectrums presented in this section are estimated with the method of [Barral and Gonçalves, 2009].

TCP performance parameter	associated function Φ
Individual throughput of flow f	$\Phi(X_i) = X_i^{(f)},\ f = 1, 2$
Total throughput	$\Phi(X_i) = X_i^{(\text{sum})} = X_i^{(1)} + X_i^{(2)}$
Fairness	$\Phi(X_i) = X_i^{(\text{diff})} = X_i^{(1)} - X_i^{(2)}$

Table 8.1: Different TCP performance parameters, and associated function Φ in the case of two competing TCP connections. The state $X_i = \left(X_i^{(1)}, X_i^{(2)}\right)$ represents the congestion window of the connections 1 and 2 at time i (in RTT unit).

Experimental setting

In all of our experiments, the topology of Figure 4.3 is used between the sites of Lyon and Rennes (RTT = 12 ms) with no 10 Gbps cross traffic at the principal router of Lyon. The bottleneck is at the Lyon switch, of capacity 1 Gbps. The size of the limiting buffer is 96 packets, and corresponds to a 1.2 ms maximal queueing delay which does not significantly modify the RTT value.

Long-lived TCP connections share the bottleneck with two possible types of UDP cross traffic:

- constant cross traffic, generated by a single UDP source at 835 Mbps,
- random cross traffic generated by 32 ON/OFF UDP sources, with mean OFF times and mean ON periods of 0.12 s and 0.1 s respectively. Each source's rate is limited at 40 Mbps, leading to a mean aggregate traffic load of 580 Mbps. The ON and OFF distributions are chosen exponential, so as not to generate long-range dependent aggregate traffic.

The experiment's duration is set to 1 hour, during which the TCP parameters (congestion window and detected losses) are collected every 3 ms via web100 reports. To avoid transient behaviors, we discard the first 60 s, and we retrieve from the remainder the congestion window size's time series resampled at the RTT rate. Results presented in the following stem from the analysis of these processes.

Except in Section 8.2.2, all of our results describe TCP connections using Reno variant.

One single TCP flow

Deterministic losses

In the case of a single TCP connection sharing the bottleneck with constant UDP traffic, the TCP connection deterministically experiences losses every time its congestion window has reached a maximal value compatible with the available bandwidth, here 147 packets. The congestion window oscillates between 73 and 147 packets and its mean in sufficiently large time windows should then constantly be equal to $(147 + 72)/2 = 110$, leading to a degenerate theoretical spectrum displayed on Figure 8.1. A degenerate spectrum is a spectrum reduced one single point. It is characteristic of a monofractal scaling law, where the scale-invariant quantity can take only one value. Here, this absence of variability is due to the determinism of the congestion window evolution, implying, for the mean congestion window in sufficiently large time windows, a unique possible value. Figure 8.1 shows that

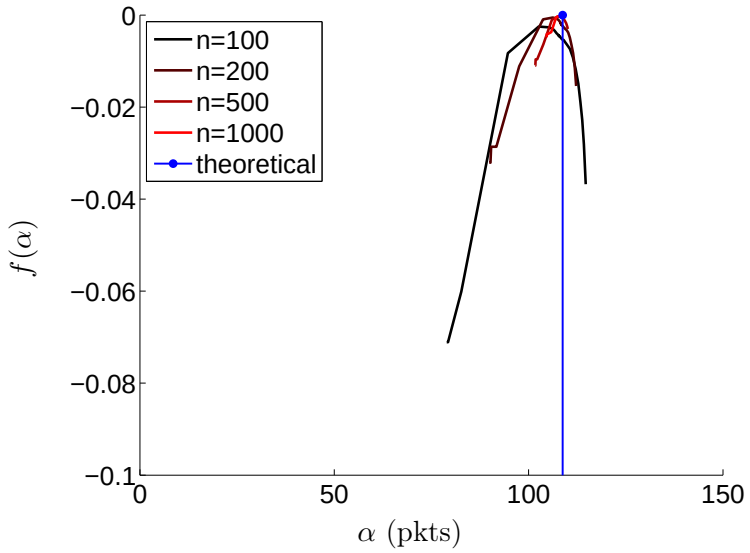

Figure 8.1: One TCP connection / constant UDP cross traffic. Theoretical (blue) and empirical (black to red) large-deviation spectrum estimated at different scales: $n = 100, 200, 500$ and 1000.

the empirical spectrum indeed converges toward the theoretical spectrum. The non-zero width of the empirical spectrum is due to a non-integer number of periods of the congestion window evolution within a time window of size n, and is naturally decreasing with n (as the number of such periods increases).

Random losses

As a realistic example of random experimental conditions that significantly differs from the commonly used Bernoulli assumption, we choose to generate ON/OFF cross traffic as described in Section 8.2.2. In Figure 8.2, statistical characteristics of the resulting traffic aggregated at buffer scale (1.2 ms) show a normal distribution (due to a sufficient number of sources), and a negligible correlation beyond the mean ON period (0.1 s). Note that the ON and OFF distributions have been chosen exponential, so that the cross-traffic does not possess long-range dependence. Thus, the loss probabilities does not depend on the past losses and satisfy the assumptions made in Section 8.2.1. Even though long-range dependent cross-traffic is likely to have minor impact on the multifractal properties investigated in this chapter if the scales involved are clearly separated, a complete study of this case is beyond the scope of our simple example aimed at illustrating the method.

As it is a basic component of our model, we now specify the loss probability as a function of the congestion window (the function $p(\cdot)$ of Section 8.2.1). A first possibility is to use directly the empirical observation, displayed in Figure 8.3. However, to permit numerical studies in cases where this empirical observation is not available, we also propose the following model. Assuming that TCP packets are regularly spaced within a RTT, the bandwidth of the TCP connection during the RTT is then constant, equal to $\mu_{\mathrm{TCP}}(cwnd) = \frac{cwnd \cdot pktsize}{\mathrm{RTT}}$ in bps. Total traffic inherits Gaussian characteristics from UDP cross-traffic with the mean shifted by $\mu_{\mathrm{TCP}}(cwnd)$. Loss probability is then simply estimated as the

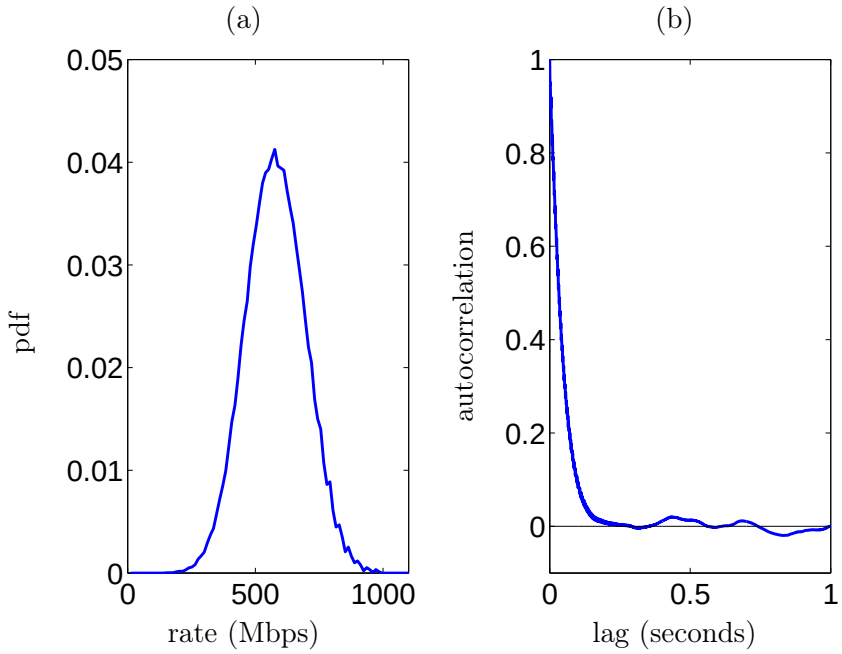

Figure 8.2: Statistical characteristics of the ON/OFF UDP cross traffic aggregated at buffer scale (1.2 ms). (a) Probability density function estimation – (b) Autocorrelation function estimation.

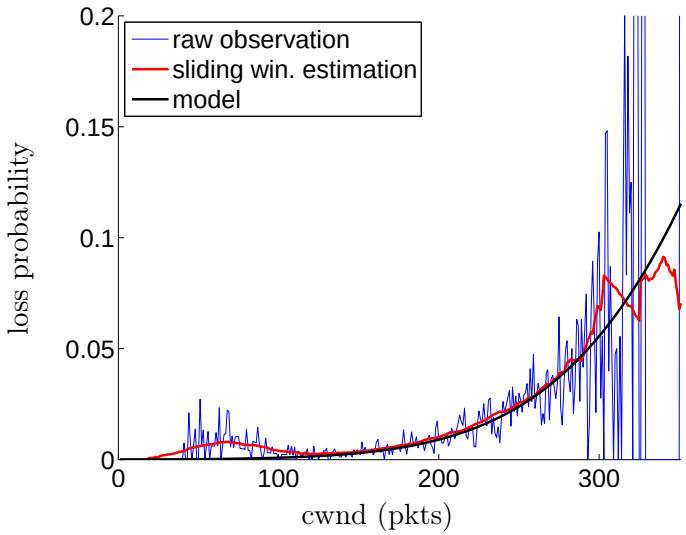

Figure 8.3: One TCP connection / ON/OFF UDP cross traffic. Loss probability as a function of the congestion window: (blue) raw empirical observation – (red) sliding window empirical estimation (window size: 20) – (black) model of equation (8.45), with estimated value $\sigma = 108$ Mbps.

probability that the total traffic aggregated at buffer scale exceeds the link capacity C:

$$p(cwnd) = 1 - \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{C - (\mu_{\mathrm{TCP}}(cwnd) + \mu_{\mathrm{UDP}})}{\sigma\sqrt{2}}\right)\right), \tag{8.45}$$

where σ is the variance of the UDP cross traffic at buffer scale and erf denotes the error function arising from the cumulative Gaussian distribution. Note that this reasoning can easily adapt to any distribution, and is not limited to the Gaussian case.

Figure 8.3 compares the model of equation (8.45) to the empirically observed loss probability and shows a remarkably good agreement for large congestion windows (roughly beyond 100). Such a good agreement is likely to be contextual: Firstly, in our real environment, we observed that packets do not arrive in burst at the beginning of an RTT, but are actually spanned over the RTT (with a slight peak at the beginning), for example due to random usage of the machines' CPU. Secondly, queueing effects that are not taken into account in equation (8.45) are actually negligible in our experimental setting because of the small buffer size. The slight discrepancy observed between the model and the data for smaller windows (less than 100 pkts) is due to the non-zero correlation function of the cross traffic for lags up to roughly 100 RTT (120 ms). This persistence effect promotes consecutive packet losses and thus accentuates the probability of TCP loss at rather small congestion windows. Note that this non-Markovian effect, which remains limited in our experiments, could limit the applicability of our simple Markov chain based model in case of an ON/OFF cross traffic with very large ON period mean.

Returning to the congestion window size's time series, we estimate the empirical large-deviation spectrum as defined in (8.38), using the method proposed in [Barral and Gonçalves, 2009]. Then, Figure 8.4 shows a clear superimposition of the spectra obtained at different time scales (from 100 to 1000 RTT). This scale-invariance property is reminiscent of the presence in TCP traffic of the multifractal scaling laws we devised in the previous section. To verify its origin, we compare these empirical estimates to the theoretical spectrum expression of a Markov chain, where the modeled loss probability density (8.45) or its empirical estimate can be used indifferently. In both cases, we obtain good agreement. It confirms the ability of our simple Markov chain model to account for the essential characteristics of TCP control mechanisms which generate the multifractal structure observable on a real TCP implementation trace. The difference observed at small α's between the two theoretical predictions and the estimated spectrum, comes from the loss probability overshoot for small window sizes. The loss model of equation (8.45) does not take it into account and therefore leads to an underestimated number of intervals with small mean throughputs. On the other side, this overshoot leads to overestimating the number of intervals where the congestion window remains small when using the raw empirical loss probability whereas in reality, consecutive losses periods responsible for this overshoot are followed by periods without loss increasing the congestion window (recall the persistence effect of the cross traffic). The good global match between empirical spectra and the theoretical spectrum using the raw empirical loss probability confirms that the non-Markovianity has very slight consequences in our experiment.

Remarkably though, notice the good match between empirical and theoretical values of the bounds $\alpha_{\min}$ and $\alpha_{\max}$ of equation (8.41), which complete the scaling-law description and allow us to accurately predict the minimal and maximal values of the mean throughput expectable at a given scale in a finite-duration experiment.

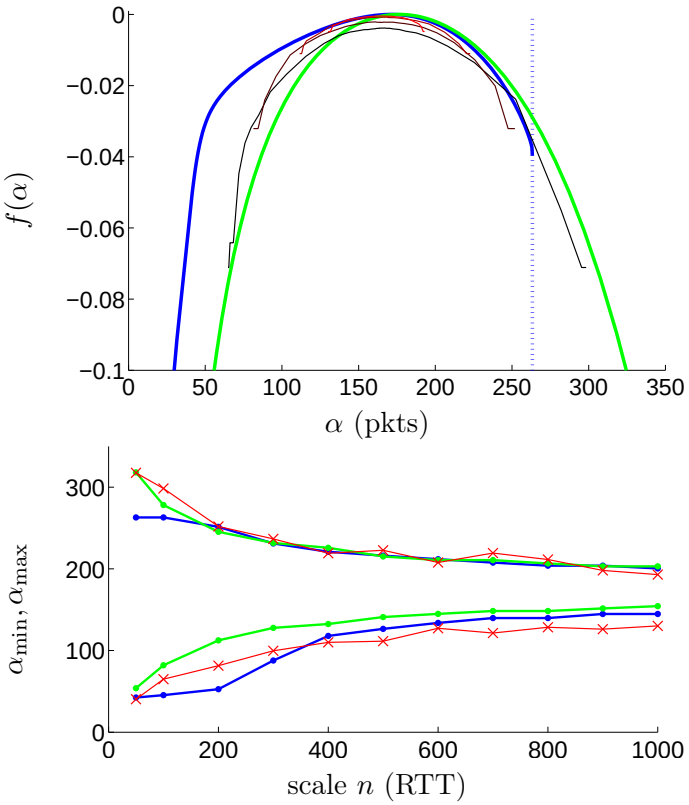

Figure 8.4: One TCP connection / ON/OFF UDP cross traffic. (top) large-deviation spectra: (black to red) empirical at scales $n = 100, 200, 500, 1000$; (blue) theoretical using raw empirical losses; (green) theoretical using the model of equation (8.45) – (bottom) min and max bounds of the spectrum: (red) empirical; (same colors as for the spectrum) theoretical.

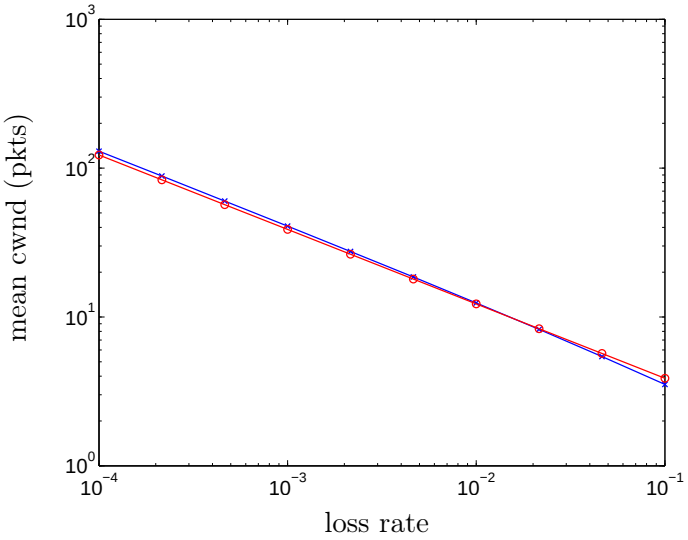

Figure 8.5: One TCP connection / Bernoulli losses (numerical simulations): comparison with the square root formula of [Padhye et al., 1998]. (blue) mean congestion window obtained as the apex of the theoretical large-deviation spectrum – (red) Mean congestion window from the square root formulae.

As compared to the deterministic case of Figure 8.1, the spectrum of Figure 8.4 corresponding to the random-loss case have a large width, characteristic of a high variability of the mean throughput in time windows of size n. Even though the observed range of variability (reflected by the bounds $\alpha_{\min}$ and $\alpha_{\max}$) naturally decreases with the averaging period n, it remains significant over 1000 RTTs (between 125 and 190 pkts); and the values of the spectrum close to zero denote a rather large probability to observe a mean throughput in a time window of size n deviating from the almost-sure "global mean" corresponding to the apex of the spectrum. Note finally that although the theoretical result of equation (8.38) rigorously holds in the limit $n \to \infty$, Figure 8.1 shows that it approximately holds above $n = 100$ which corresponds to scales of interest in network applications.

Bernoulli losses

We finally mention that in the case of the widely used assumption of Bernoulli losses, our model is fully consistent with previously derived results. In particular, Figure 8.5 shows that the apex of the corresponding large-deviation spectrum coincides with the mean throughtput predicted by the well-known square root formula of [Padhye et al., 1998] (the one which does not include timeout effects, see equation (3.5) of Chapter 3). This coherence with simpler models considering only the mean throughput is an important aspect of our model, which can also be used to obtain first order statistics of diverse TCP performance parameters in various conditions.

Two competing TCP flows

We now study the case of two competing TCP connections in constant and random cross traffic. To handle the case of two connections, we use the model described in Section 8.2.1

for vectorial-states Markov chains. We use the assumption of Section 8.2.1 according to which the loss probability depends only on the sum of the congestion windows. We use the empirical loss probability. For constant cross traffic, it is almost a step function taking value 1 for congestion windows larger than 150. For the ON/OFF cross traffic, it is similar to Figure 8.3.

Steady-state distributions

As a first assessment of the ability of the model presented in Section 8.2.1 (in particular equation (8.44)) to reproduce basic properties of TCP traffic, Figure 8.6 compares the theoretical and empirical steady-state distributions in both cross-traffic conditions. Synchronization rates used in the blue curves of this figure are estimated from the traces (syn = 0.57 for the constant cross traffic and syn = 0.85 for the ON/OFF cross traffic). Mismatch between the theoretical curves when the syn parameter is assumed equal to zero, and the empirical curves confirms that synchronization is an important parameter of the system. The quasi-superimposition of the curves with the estimated synchronization rate and with the joint probability shows that the effect of synchronization is well taken into account with our single parameter syn. Finally, good match of these two curves with the empirical data supports our assumption that the loss probability is only a function of the sum of the congestion windows.

Besides validating our model, Figure 8.6 also confirms natural intuitions on the effect of synchronization, mainly that it tends to equalize the congestion windows of the two flows, and then to reduce their variability and the variability of their difference. It also naturally reduces the mean of their sum (and thus the global performance).

Spectrums

Large-deviation spectra with constant and ON/OFF cross traffic conditions are displayed in Figures 8.7 and 8.8. We first observe that these figures confirm again the presence of our novel multifractal scaling law in the case of two TCP connections as their empirical spectra overlay at different scales. Comparison of the theoretical spectra with and without synchronization confirms our previous conclusions about the synchronization effect: it tends to reduce the variability of the congestion window of individual flows, and the variability of their difference, thus reducing the mean of their sum.

However, whereas theoretical and empirical spectra of the sum fairly match, empirical spectra of the individual congestion window and of the difference are much narrower than their theoretical predictions, even with the inclusion of the synchronization parameter. Let us recall that we assume a loss probability independant of the individual windows (for a given sum). We could easily refine this hypothesis and distribute this loss probability conditionnaly to each window size. Assigning a higher loss probability to the flow with the larger congestion window certainly would reduce the spectrum's width.

This slight mismatch between our model and the experiments due to oversimplification could not be detected based on the steady-state distributions' observation. The new multifractal scaling laws that we identified in this work can then be seen as a fine analysis which can turn very useful in checking a model's adequacy to real systems.

Finally, note the very good match between theoretical and empirical bounds when synchronization is taken into account.

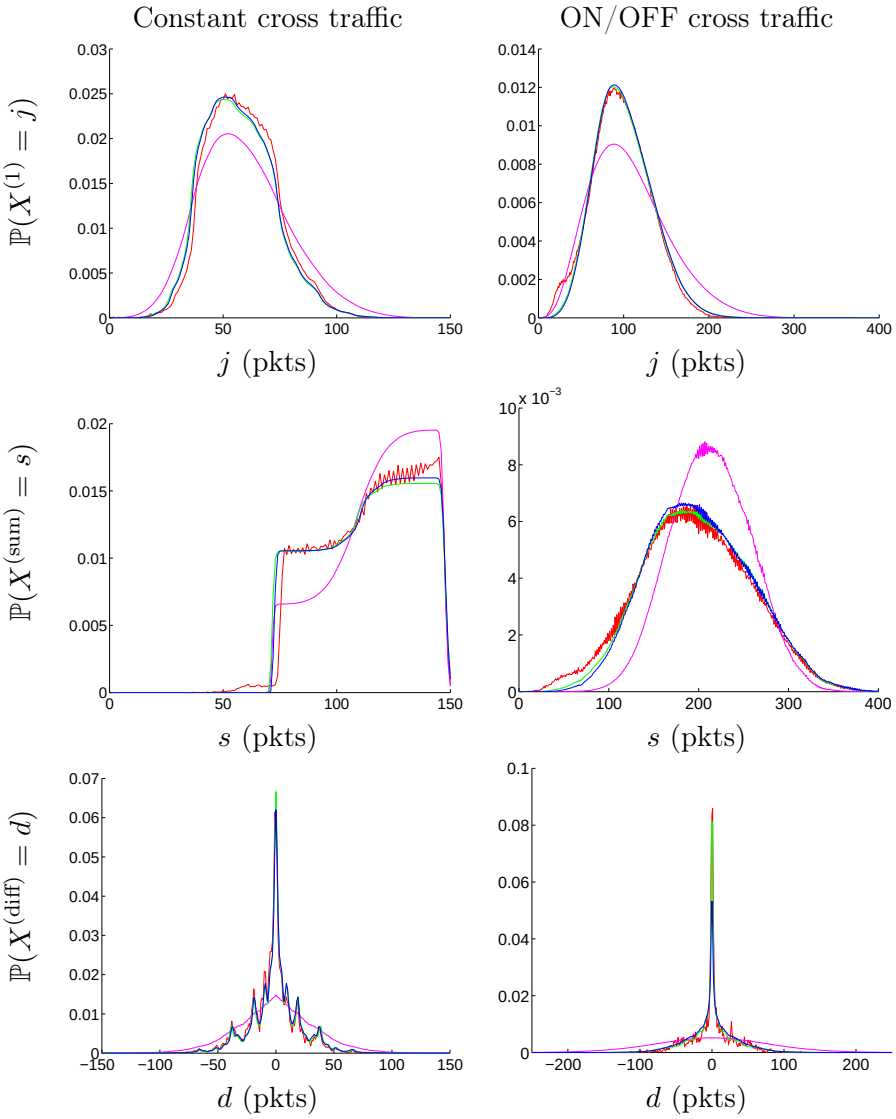

Figure 8.6: Two TCP connections. (left) Constant UDP cross traffic – (right) ON/OFF UDP cross traffic. Steady-state distributions of (top) the congestion window of flow 1 – (middle) the sum of the congestion windows – (bottom) the difference of the congestion windows. (red) empirical observation – (magenta) model of equation (8.44) with syn = 0 – (blue) model of equation (8.44) with syn = 0.57 for constant UDP cross traffic and syn = 0.85 for ON/OFF cross traffic – (green) utilization of the empirical joint loss probability.

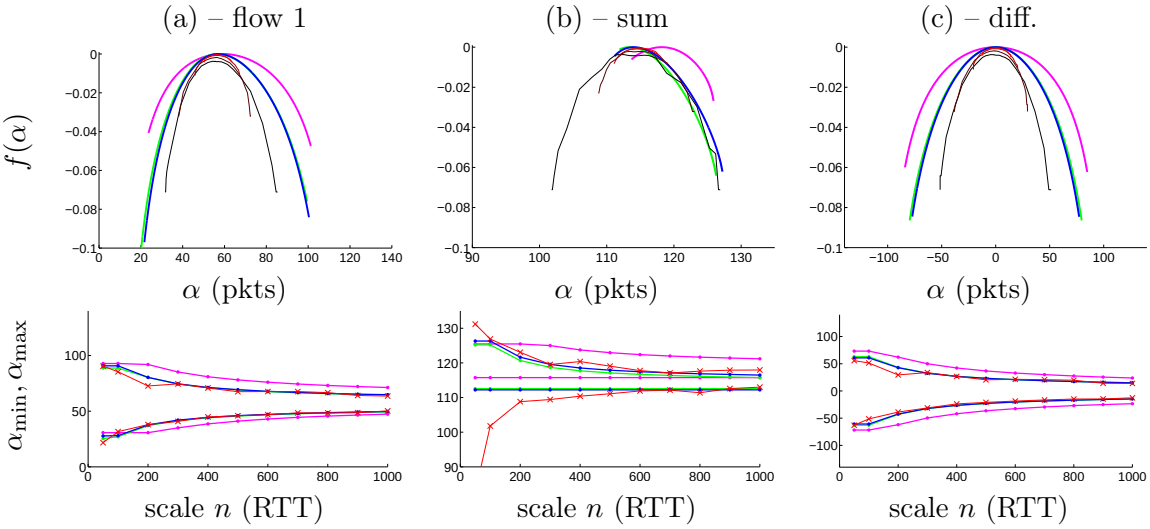

Figure 8.7: Two TCP connections / Constant cross traffic. Large-deviation spectra and bounds for (a) the congestion window of flow 1 – (b) the sum of the congestion windows – (c) the difference of the congestion windows. (top) large-deviation spectra: (black to red) empirical at scales $n = 100, 200, 500, 1000$; (magenta) theoretical with syn $= 0$; (blue) theoretical with syn $= 0.57$; (green) theoretical with the empirical joint loss probability – (bottom) min and max bounds of the spectrum: (red) empirical; (same colors as for the spectrum) theoretical.

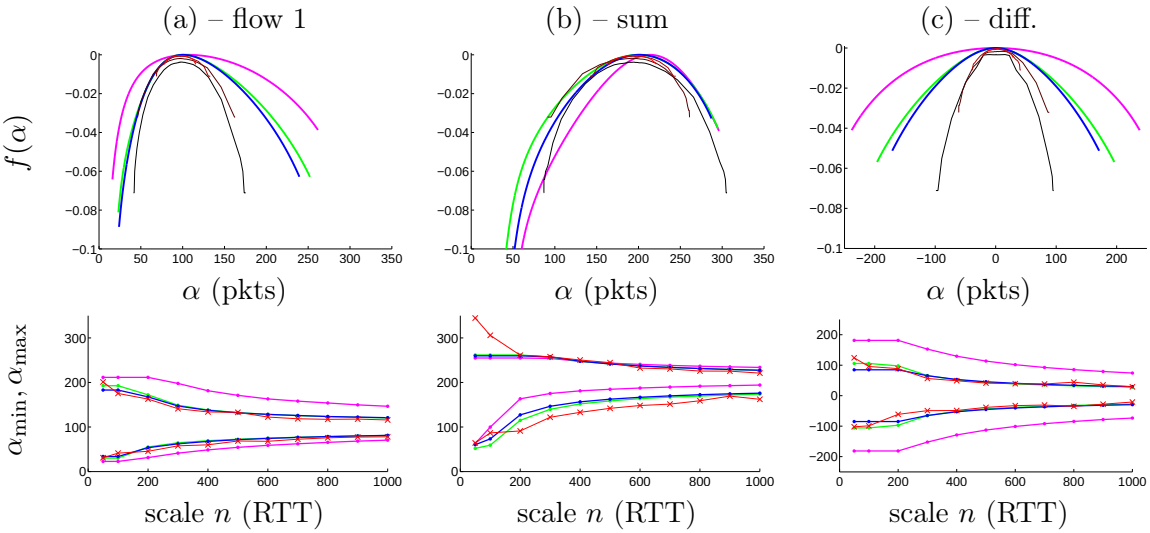

Figure 8.8: Two TCP connections / ON/OFF cross traffic. See legend of Figure 8.7, blue curves correspond here to syn $= 0.85$.

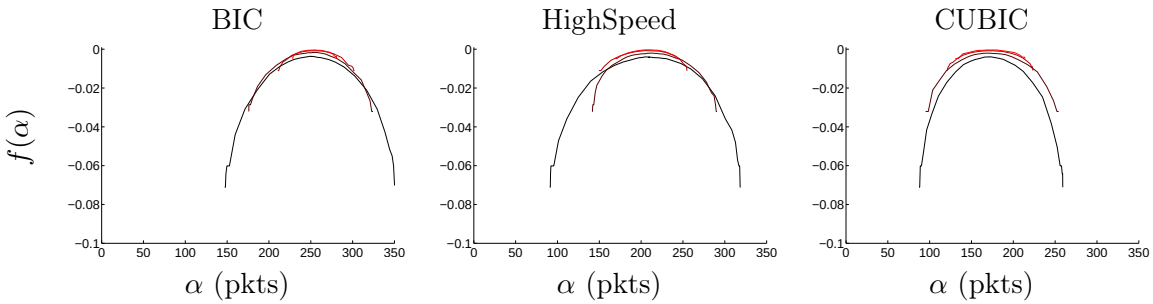

Figure 8.9: One TCP connections / ON/OFF cross traffic. Large-deviation spectra. Empirical large-deviation spectra at scales $n = 100, 200, 500, 1000$ for three other TCP variants than Reno.

Different TCP variants

As examples of other TCP variants, Figure 8.9 shows the spectra of a single connection with ON/OFF cross traffic. The TCP variants presented are BIC, HighSpeed and CUBIC (compare also with Figure 8.4 for the case of Reno). Even though further work extending the model of Section 8.2.1 is required to have theoretical predictions of these spectra, we conclude from the good superimposition of the spectra at different scales observed on this figure, that other TCP variants than Reno also inherently possess scale-invariance properties of the same kind. Even though the shape of the spectra are different for the different variants, the most important difference is certainly their apex, which reflect the variant's aggressivity. Note that Figure 8.10 shows that the model proposed in the previous section for the loss probability as a function of the congestion window remains quite accurate with these other TCP variants.

As examples of the case of two TCP connections including other variants than Reno, Figure 8.11 shows the spectra corresponding to the case where one TCP Reno connection is competing with a BIC or a CUBIC connection, with constant cross traffic. This figure confirms the scale invariance property of the large-deviation spectrum corresponding to different TCP performance parameters, and highlight its possible use to perform a *multi-scale analysis* of the "TCP friendliness" of other TCP variants than Reno: the repartition of the mean fairness between several competing TCP flows of different variants, in time windows of size n, is invariant across scales.

8.2.3 Conclusion

In this section, we demonstrated that real TCP traffic, generated on physical machines over a real network, inherently possess a new type scaling laws directly related to TCP performance parameters. The scale-invariant quantity is the averaged throughput or fairness in time windows of size n, and it exhibits a multifractal behavior.

We proposed a Markov chain model of the TCP congestion window evolution based on the AIMD mechanism. Then, using the large-deviation theorem that we provided in previous section, we showed that our model accurately reproduces the large-deviation spectrum in several examples of experimental conditions. This conclusively assesses the fundamental role of the AIMD mechanism in generating the newly defined scaling laws.

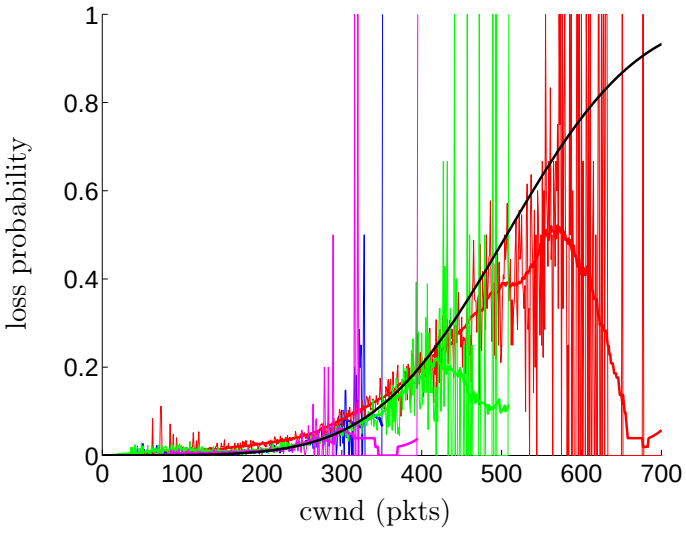

Figure 8.10: One TCP connection / ON/OFF UDP cross traffic. Loss probability as a function of the congestion window: (blue) Reno – (red) BIC – (green) HighSpeed – (magenta) CUBIC – (black) model of equation (8.45), with estimated value $\sigma = 108$ Mbps. The continuous curves represent the sliding window estimations.

The main strength of our model is that it can include several competing TCP connections, and take into account synchronization effects between these connections.

Although the scaling law theoretically holds for very large scales only, our experimental results show that it holds for scales larger than $n = 100$ RTT, which correspond to scales of interest in applications. Moreover, the bounds on the observable spectrum on finite-length traces are also well predicted, which completes the description of the spectra and also gives valuable practical information for applications.

Our experimental results show that similar scaling laws are also present in TCP traffic using other variants than Reno. Inclusion of these variants in our analytical model would then be an interesting direction of future work, to offer the possibility of an analytical *multi-scale analysis* of the fairness between several connections using different variants, directly related to the "TCP friendliness" of these variants.

Conclusion

In this chapter, we have shown that TCP traffic possess new types of multifractal scaling laws, whose scale-invariant quantities correspond to average TCP performance parameters such as throughput or fairness, in time windows of a certain size n.

We first provided an almost-sure large-deviation theorem, in a general mathematical context. The originality of this theorem lies that it considers the empirical mean of a process on one realization, instead of its ensemble mean. It then theoretically assesses the presence of the new scaling laws on one realization of the process, but also allows us to predict bounds on the corresponding large-deviation spectrum observable on a finite-size realization.

Application of this theorem to a Markov chain model of the AIMD mechanism of TCP

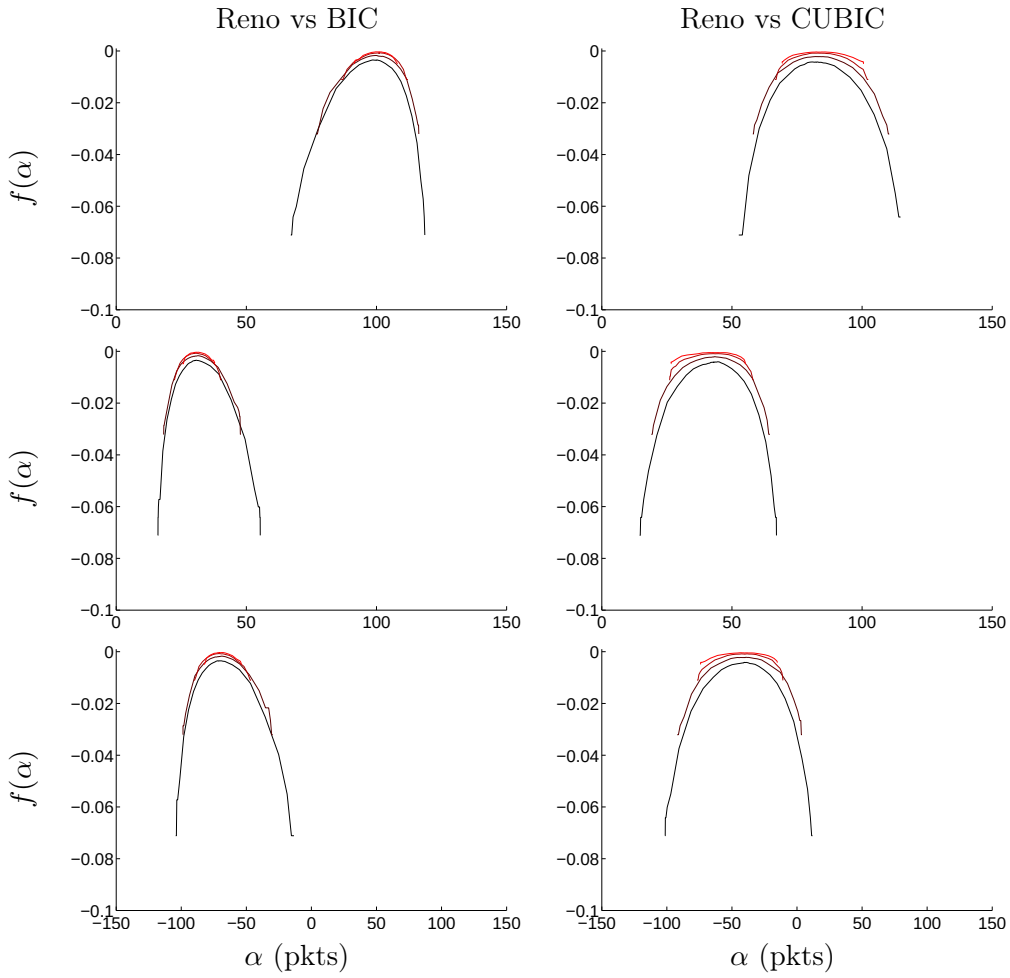

Figure 8.11: Two TCP connections / Constant UDP cross traffic: case of one Reno connection competing with one connection using another variant: (left) BIC – (right) CUBIC. Large-deviation spectra at scales $n = 100, 200, 500, 1000$ for the congestion windows: (top) individual congestion window for the connection using the other variant – (middle) individual congestion window for the Reno connection – (bottom) fairness $(X^{(1)} - X^{(2)})$ where $X^{(1)}$ is the congestion window of the Reno connection and $X^{(2)}$ the congestion window of the other connection.

allowed us to show that this mechanism is genuinely responsible for the presence of the new scaling laws identified in TCP traffic, and to compute the corresponding theoretical large-deviation spectrum. Our model is able to include several competing TCP connections (potentially synchronized), and provides a support to perform an analytical multi-scale analysis of TCP performance parameters throughput and fairness. Such an analysis permits the prediction of probabilistic bounds on this performance parameters of primary interest for QoS issues. Moreover, our model can easily include various types of evolution of the TCP congestion window, and it can then be used to find the optimal evolution.

This work has been done in the context of long-lived TCP connections and it can be extended in several directions. Firstly, our model assumes stationary loss probabilities which do not depend on past losses. Relaxation of this assumption, for example using inhomogeneous Markov chains, would be an interesting direction to include more realistic scenarios than the examples presented in this chapter. In the same lines, using heterogeneous sources, and including variable RTTs and queueing delays, would also permit to handle more realistic situations. Inclusion of the ON/OFF structure would be another interesting direction to embed in a single framework long-range dependence due to heavy-tailed flow size distributions and multifractality due to the AIMD mechanism. Also, theoretical spectra are obtained through the computation of spectral radii, which can have high computational costs with large matrices corresponding to a large number of competing flows, thus limiting the scalability of our method. Given the potential applications in optimization procedures, design of efficient procedures in this case could give our method higher practical capabilities of primary interest.

TCP traffic at packet scale and TCP throughput predictability.

- Do TCP mechanisms generate multifractal scaling laws in the packet-level source traffic? What is their relation with TCP throughput prediction (beyond the mean)?

The AIMD mechanism generates multifractal scaling laws in TCP traffic from one connection. The scale-invariant quantity is the averaged throughput in time windows of size n. A Markov model allows us to demonstrate this multifractal property and to calculate the corresponding theoretical large-deviation spectrum observable on one realization based on an almost-sure large-deviation theorem. Bounds on the spectrum for a finite-size realizatioin can also be predicted.

In the case of multiple connections, these scaling laws are also present and permit a multi-scale analysis of TCP performance metrics such as the total or individual throughput or fairness.

The large-deviation spectrum permits the prediction of probabilistic bounds on the throughput and fairness of TCP connections.

Chapter 9

Conclusions and perspectives

Quality of Service is a central preoccupation for network users, intimately linked to the properties of the traffic. While several approaches are possible, we adopted in this thesis a probabilistic approach, based on a deep understanding of the statistical properties of network traffic and of their impact on QoS.

Conclusions

A thorough review of the state of the art regarding scaling laws in network traffic and their impact on QoS (Chapter 3) revealed that these questions have been at the center of a large amount of research works over the last decade. However, such works belong mainly to two categories, theoretical or experimental, which are sometimes separated by such a gap that theoretical models can fail at describing the results of real-world experimentations. While this is a common fact in many fields of science, in the case of networks, we can identify a major cause for this division: the TCP protocol. Despite its apparent simplicity (in its basic variant, Reno), the TCP protocol introduces a large complexity in the system. Indeed, its feedback mechanisms inevitably introduce correlations between sources sharing a bottleneck. Taking this correlation into account in mathematical models is a very arduous challenge, far from being solved at this time. In this sense, experimental studies are essential to provide accurate results in specific conditions. On the other hand, our state-of-the-art review also evidenced the risks of drawing general conclusions from specific experimental results, and more particularly from the analysis of a specific Internet trace. While these studies are essential to understand the main evolutions of Internet traffic, such conclusions are likely to be trace-driven and not to hold in full generality. In this sense, theoretical models can help to understand the genuine origin of experimental results, and to determine their limits and conditions of validity in different situations. In our work, we have tried to keep a constant interaction between these theoretical and experimental aspects.

To performed controlled, yet realistic experiments, we developed a large-scale experimental platform (Chapter 4). Based on the testbed *Grid5000*, this platform consists in a large number of fully controllable machines interconnected by very high-speed dedicated links, and allows reproducing realistic conditions with real network equipments and TCP implementations. We also developed a tool, *MetroFlux*, able to capture traffic at packet level, at up to 10 Gbps. This tool also enables synchronized captures at the input and the output of a buffer, to study its dynamic. It then permits a deep inspection of both traffic and QoS properties. Despite the natural difficulty inherent to real equipments shared by a large research community, we managed to obtain a reliable and powerful metrology

platform which we used throughout the rest of our work.

This metrology platform first allowed us to perform a thorough investigation of Taqqu's relation between heavy-tails and long-range dependence (Chapter 5). Then, we could close a ten-year-old debate about the role of the TCP protocol in generating long-range dependence. We showed that, because of the scale range involved in Taqqu's relation, the protocol is, by itself, unable to generate such scaling behavior in loss-free situations. Moving to more realistic conditions implying losses, we showed that the interaction of the protocol and the system can modify the observed long-range dependence. In particular, in congestion situations, RTT-scale correlations introduced by TCP feedback mechanisms can have an impact even on large-scale traffic properties. This controlled experimental study allowed us to clarify the situations in which simplifications of Taqqu's model render it incompatible with real traffic properties. Then, to include a commonly observed situation where the flow-size and flow-duration distributions have different tail indices, we proposed an extension of Taqqu's model taking into account the flow-scale correlation between rates and durations. This model is based on a planar Poisson process, a setting rarely used in the context of long-range dependence, which proved well suited to predict the Hurst parameter of the aggregate traffic. Our results show that, depending on the precise form of this correlation, surprising effects can appear: the long-range dependence can be even stronger than the one predicted by the original Taqqu's relation.

The study of long-range dependence confirms that the tail index of the flow-size distribution is an important parameter of the system. While its estimation from packet traces enters a well-known framework if all the packets are observed, the situation is more complicated when resource-usage limitations constrain to retain only a sub-sample of the packet stream. We proposed a rigorous maximum-likelihood solution and showed that other proposed solutions are approximations of this estimator (Chapter 6). Practically, we showed that it is the only reliable estimator at sampling rates as small as 1/1000, advocating its use in long-duration monitoring of very high-speed traffic.

Turning to the evaluation of QoS and its link with scaling laws, we used two different approaches. We first performed an empirical evaluation of the impact of the flow-size distribution's tail index α_{SI} on QoS (Chapter 7). With the UDP protocol, we observed a clear deterioration of QoS for very heavy tails and for large buffers. In contrast, we showed the insensitivity of QoS with respect to α_{SI} for small buffers. These results are coherent with asymptotic theoretical results proposed in the literature. However, our empirical approach allowed us to go beyond asymptotic results and to specify the bounds between the large-buffer and the small-buffer regimes. In particular, we observed different regime frontiers for the loss rate and for the mean buffer occupancy, stemming from the different statistical nature of these quantities. In the case of TCP, the impact of α_{SI} on QoS was less clear. Our results suggest that the absolute position of the whole distribution might be as important as its tail index for the evaluation of QoS. However, our experimental results showed evolutions of certain QoS metrics with α_{SI} in contradiction with several prior results of the literature. This tends to prove that the results in TCP can strongly depend on the experimental conditions and advocates the use of real experiments as a good approach to tackle QoS issues in TCP.

As a complementary aspect to the flow-scale characterization of QoS, we finally turned to the packet-scale characterization of TCP traffic within a long flow (Chapter 8). Based on an original large-deviation theorem, we showed that the TCP congestion window's evolution exhibits a new type of mutlifractal scaling laws, genuinely originating in the AIMD

mechanism. The scale-invariant quantity is the mean TCP throughput within some time window, and the corresponding large-deviation spectrum thus directly reflects a multi-scale analysis of TCP performance. We proposed a Markov chain model of the congestion window's evolution, which permits to simply compute the theoretical spectrum and to predict probabilistic bounds on the expected throughput. Such bounds, accounting for the traffic variability, enable a finer description than the predictions of the global mean throughput existing in the literature. Beyond the versatility of the proposed model with respect to experimental conditions, its most important strength lies in that, using higher dimension Markov chains, it can take into account the RTT-scale correlation between competing sources. Then, in the case of several flows, our model also allows performing a multi-scale analysis of fairness and it gives the corresponding QoS bounds. Finally, we experimentally showed that other TCP variants than Reno also inherently possess the same kind of scaling laws.
Let us highlight that our multifractal approach is different from the standard Hölder multifractality, which, in spite of many efforts, was never proven to straightforwardly connect to TCP mechanisms. In this sense, our work offers a fresh view on the widely debated question of the link between TCP and multifractality, and opens new perspectives.

Perspectives

The Internet is currently undergoing deep modifications both in term of structure, with the emergence of FTTH and flow-aware control mechanisms; and in term of users' behavior, with the development of new aggressive TCP variants and the expansion of UDP-based traffic. A precise understanding of the impact of these new mechanisms on traffic and QoS is essential to ensure the best performance and stability for the future Internet. Such a deep understanding, in our opinion, can only be based on a combination of theoretical studies and experimental investigations. In this context, our work opens new perspectives of future work on both aspects.

The metrology platform that we developed will remain useful to many network researchers, to perform experimental studies implying the new mechanisms of the future Internet. Indeed, its high-speed links allows reproducing a large spectrum of rates, such as the ones that will be available at home in the future Internet, with the emergence of FTTH. Moreover, although our work in this thesis used only a nation-wide network, *Grid5000* also offers several international connections (with Japan, Brazil and the Netherlands). These long-distance connections can be used to perform experiments with heterogeneous and/or large RTTs, *e.g.*, to study the behavior of new aggressive TCP variants developed for large bandwidth-delay products. The deployment of *MetroFlux* in Japan is already planed, via the ANR[1] project PetaFlow[2], and will permit a deep analysis of the traffic properties in such situations.

Regarding the characterization of the aggregate traffic properties, some parts of the problem have been left over. We saw that the RTT-scale correlations play an important role, but our analysis remained qualitative on this point. A quantitative study of their precise impact and of the scales involved is then a natural future work. Extension of our study to situations implying multiple RTTs, and multiple bottlenecks would be an inter-

[1] Agence National pour la Recherche, (French national research agency)
[2] ANR Strategic Japanese-French Cooperative Program on "Information and Communications Technology Including Computer Science"

esting complementary direction.
The model that we proposed to extend Taqqu's relation also opens up several directions for further work. From a theoretical viewpoint, a deeper examination of the processes would be required to prove their convergence properties in a more general setting. It is likely that, as for the classical models which do not include the correlation between flow rates and durations, there exist (at least) two limiting regimes yielding different processes (like the fBM and the Lévy process in Taqqu's model). Several hints lead us to believe that the introduction of this correlation might also lead to interesting scaling properties of aggregate traffic at small scales, of great relevance in a network context. From an application viewpoint, the question of the precise origin of the flow-scale correlation between rates and durations remains open, but answers to this question should probably be sought in a more general context including the effects of flow-aware control procedures, susceptible to modify strongly these flow-scale correlations. Then the model we proposed could enable an analysis of the impact of these new mechanisms on the aggregate traffic properties.
Finally, to complete the description of large-scale properties of aggregate traffic, a model embedding both the RTT-scale correlation (introduced by TCP) and the flow-scale correlation (possibly introduced by flow-aware control procedures) would be of great interest. Such a model would for example allow studying the interactions between the TCP protocol and the flow-scale control procedure, through their effect on the aggregate traffic.

Our study of large-scale properties of the traffic indicates that the tail index of the flow-duration distribution can be important if it is different from the one of the flow-size distribution. A promising direction to extend our work on the estimation of the flow-size distribution's tail index under sampling would then be to adapt it to the case of the flow-duration distribution. In order to use the tail parameter in control or optimization procedures, real-time constraints require extra work to reach a fully operational estimator in term of computational cost. Finally, this estimator has been developed within the framework of Pareto distributions and its adaptation to more general heavy-tailed distributions could also be an interesting direction to extend our work.

The empirical approach that we used to evaluate the impact of the flow-size distribution's tail index on QoS allowed us to observe some new results on queueing behaviors, but many questions remain open on this aspects. Our empirical results on loss rate and buffer occupancy in UDP are intimately related to the finite-size character of the buffer, for which mathematical modeling is very arduous. It then exposes a challenging direction for future theoretical work. Such work on QoS in UDP is likely to have a major importance in the future Internet with the expansion of UDP-based applications. Moreover, our study was limited to Gaussian input traffic, and a natural extension would be to consider more general input processes. Given the non-linear queueing effects, the situation is likely to turn much more complicated. Such a work, however, could be of major importance, because non-Gaussian traffic is likely to appear more prominently in the future Internet, due to the emergence of FTTH, increasing the probability of high-rate bursts.
In the case of TCP, the tail index did not appear as a sufficient parameter of the flow size distribution to characterize QoS, and this result offers challenging future work to find more pertinent parameters of the distribution. As we saw, the results are also likely to be sensitive to some complex optimization parameters of TCP, such as the `SSTHRESHOLD`. A precise characterization of the impact of these parameters is then essential, and the use of our controlled metrology platform will be of great help is this quest. The extension of our study to different TCP variants is also an aspect that we have left over, and that would

be an important future work.

The multifractal approach that we proposed to characterize TCP traffic at packet-scale also opens new directions of future work, both in an network application context, and in a more theoretical context. Firstly, extension of the Markov model we proposed to new aggressive TCP variants would be an interesting perspective, for it would permit an analytical mulit-scale evaluation of their "TCP friendliness". Inclusion in the model of heterogeneous RTTs and queueing delays would also be an interesting direction to handle more general situations. Then, the model could be used, not only to characterize and predict achieved QoS, but also for control and optimization problems related to the TCP protocol such as finding the optimal evolution of the congestion window to maximize the performance bounds predicted by the large-deviation spectrum. To consider such approaches in real-time, however, further work would be required to enhance the practical capabilities of our approach based on the computation of spectral radii of large matrices.
The next step toward the understanding of scaling laws and QoS in networks is probably to merge the flow-scale and packet-scale approaches, *i.e.*, include the ON/OFF structure in the Markov-chain model. This is a very arduous challenge, for it implies correlations of different types, at both small and large scales.
The large-deviation theorem that we provided to show that a multifractal behavior is observable on one realization of a Markov chain also opens several directions of future works of diverse nature. Firstly, the statement of the theorem itself can be improved, in particular by finding a necessary condition for its result to hold (we used a sufficient condition which can be too restrictive). From a theoretical viewpoint, this theorem shows how to generate a new class of mutlifractal processes with prescribed large-deviation spectra. This can be of interest for the mutlifractal community, with a goal of finding new ways (not based on multiplicative cascades) to generate processes with controlled mutlifractality. Finally, we believe that this theorem can find applications in many domains other than networks, where the same kind of scaling laws is of primary interest.

In this thesis, we have tried to provide a better understanding of network traffic properties and their impact on QoS, in order to eventually permit the development of more efficient control and optimization procedures for the future Internet. Trying to draw bridges between theory and experiments, we have seen that one of the main difficulties in the mathematical description of real-world phenomena is the inclusion of different types of correlations, at different scales. While our work was motivated by network applications, this is a common problem in many fields of science like physics or economics (some authors, for instance, partially attribute the current financial crisis to erroneous predictions due to the use of the (uncorrelated) standard Brownian motion). We believe that the development of models including various types of correlations is a major mathematical challenge for current and future research, conditioning the development of accurate prediction and control procedures in many domains.

Appendix A

Technical derivations for Chapter 6

A.1 Formulae of the sampled flow size distribution, log-likelihood and Fisher information for an arbitrary value of $j_{\min}$

With an arbitrary value of the minimal sampled size observed $j_{\min}$, the sampled flow size distribution (eq. (6.19)) becomes when using the proper normalization (eq. (6.7)):

$$P_Y(Y = j|\alpha) = \frac{\sum_{i=j}^{\infty} B_p(i,j) i^{-(\alpha+1)}}{\sum_{j=j_{\min}}^{\infty} \sum_{l=j}^{\infty} B_p(l,j) l^{-(\alpha+1)}}, \tag{A.1}$$

The log-likelihood function can then be written:

$$\begin{aligned} \mathcal{L}(\alpha) = \quad & -n \ln \left(\sum_{j=j_{\min}}^{\infty} \sum_{l=j}^{\infty} B_p(l,j) l^{-(\alpha+1)} \right) \\ & + n \sum_{j=j_{\min}}^{\infty} \eta_j \ln \left(\sum_{i=j}^{\infty} B_p(i,j) i^{-(\alpha+1)} \right). \end{aligned} \tag{A.2}$$

Finally, by differentiation of eq. (A.2), the Fisher information for an arbitrary $j_{\min}$ is obtained:

$$\begin{aligned} \mathcal{I}(\alpha) = \quad & n \Bigg(\frac{\left(\sum_{j=j_{\min}}^{\infty} \sum_{i=j}^{\infty} \frac{(\ln i)^2 B_p(i,j)}{i^{(\alpha+1)}} \right) \left(\sum_{j=j_{\min}}^{\infty} \sum_{i=j}^{\infty} \frac{B_p(i,j)}{i^{(\alpha+1)}} \right)}{\left(\sum_{j=j_{\min}}^{\infty} \sum_{l=j}^{\infty} B_p(l,j) l^{-(\alpha+1)} \right)^2} \\ & - \frac{\left(\sum_{j=j_{\min}}^{\infty} \sum_{i=j}^{\infty} \ln i B_p(i,j) i^{-(\alpha+1)} \right)^2}{\left(\sum_{j=j_{\min}}^{\infty} \sum_{l=j}^{\infty} B_p(l,j) l^{-(\alpha+1)} \right)^2} \\ & + \mathbb{E}_Y \left\{ \left(\frac{\sum_{i=1}^{\infty} \ln i B_p(i,Y) i^{-(\alpha+1)}}{\sum_{l=1}^{\infty} B_p(l,Y) l^{-(\alpha+1)}} \right)^2 \right\} \\ & - \mathbb{E}_Y \left\{ \frac{\sum_{i=1}^{\infty} (\ln i)^2 B_p(i,Y) i^{-(\alpha+1)}}{\sum_{l=1}^{\infty} B_p(l,Y) l^{-(\alpha+1)}} \right\} \Bigg). \end{aligned} \tag{A.3}$$

A.2 Asymptotic relations between some normalization sums

In this appendix, we show a few relations between the normalization sums of the following distributions: Zeta, Pareto, distribution of $\overline{i}_{(\alpha^{ap})}$ assumed to be algebraically decreasing. The tail index of these distributions is denoted α.

Let us first stand and recall a few notations:

$$\zeta(\alpha+1, i_{\min}) = \sum_{i=0}^{\infty} (i+i_{\min})^{-(\alpha+1)} = \sum_{i=i_{\min}}^{\infty} i^{-(\alpha+1)}, \tag{A.4}$$

$$\psi(\alpha+1, x_{\min}) = \int_{x_{\min}}^{\infty} x^{-(\alpha+1)} \mathrm{d}x = \frac{x_{\min}^{-\alpha}}{\alpha}, \tag{A.5}$$

$$\chi(\alpha+1, j_{\min}) = \sum_{j=j_{\min}}^{\infty} \left(\overline{\imath}_{(\alpha^{ap})}(j)\right)^{-(\alpha+1)}, \tag{A.6}$$

where

$$\overline{\imath}_{(\alpha^{ap})}(j) = \exp\left(\frac{\sum_{i=j}^{\infty} \ln(i) B_p(i,j) i^{-(\alpha^{ap}+1)}}{\sum_{l=j}^{\infty} B_p(l,j) l^{-(\alpha^{ap}+1)}}\right). \tag{A.7}$$

Considering that $\zeta(\alpha+1, i_{\min})$ is the Riemann sum approximating the integral $\psi(\alpha+1, i_{\min})$, we can easily see that:

$$\zeta(\alpha+1, i_{\min}-1) \leq \psi(\alpha+1, i_{\min}) \leq \zeta(\alpha+1, i_{\min}). \tag{A.8}$$

As we also have:

$$\zeta(\alpha+1, i_{\min}) - \zeta(\alpha+1, i_{\min}-1) = \frac{1}{i_{\min}-1} \underset{i_{\min}\to\infty}{\longrightarrow} 0, \tag{A.9}$$

we conclude that

$$\zeta(\alpha+1, i_{\min}) \underset{i_{\min}\to\infty}{\simeq} \psi(\alpha+1, i_{\min}). \tag{A.10}$$

We give now an approximation of $\overline{\imath}_{(\alpha^{ap})}(j)$ when j is large: First note that the function $B_p(i,j) i^{-(\alpha^{ap}+1)}$ (considered as a function of i) mainly takes non null values in an interval centered on the value $i = \frac{j}{p}$ (where the function is maximal) and of width a few times $\sqrt{jpq}$. If j is large, the function $\ln(i)$ is approximately constant in this interval, equal to $\ln(\frac{j}{p})$. Then we can rewrite $\overline{\imath}_{(\alpha^{ap})}(j)$ by putting the ln function out of the summation:

$$\begin{aligned} \overline{\imath}_{(\alpha^{ap})}(j) &\simeq \exp\left(\ln(\frac{j}{p}) \frac{\sum_{i=j}^{\infty} B_p(i,j) i^{-(\alpha^{ap}+1)}}{\sum_{l=j}^{\infty} B_p(l,j) l^{-(\alpha^{ap}+1)}}\right) \\ &\simeq \frac{j}{p}. \end{aligned} \tag{A.11}$$

Note that this approximation can be used to reduce drastically the computational cost for the computation of $\overline{\imath}_{(\alpha^{ap})}(j)$ for large values of j.

Direct plug of equation (A.11) into equation (A.6) leads to the approximation

$$\begin{aligned} \chi(\alpha+1, j_{\min}) &\simeq \left(\frac{1}{p}\right)^{-(\alpha+1)} \sum_{j=j_{\min}}^{\infty} (j)^{-(\alpha+1)} \\ &\simeq \left(\frac{1}{p}\right)^{-(\alpha+1)} \zeta(\alpha+1, j_{\min}). \end{aligned} \tag{A.12}$$

Using the approximation of equation (A.10) then leads to

$$\begin{aligned}\chi(\alpha+1, j_{\min}) &\simeq \left(\frac{1}{p}\right)^{-(\alpha+1)} \frac{j_{\min}^{-\alpha}}{\alpha} \\ &\simeq \frac{1}{p}\frac{(\frac{j_{\min}}{p})^{-\alpha}}{\alpha} \\ &\simeq \frac{1}{p}\psi(\alpha+1, \frac{j_{\min}}{p}). \end{aligned} \tag{A.13}$$

As $\overline{\imath}_{(\alpha^{ap})}(j_{\min}) \simeq \frac{j_{\min}}{p}$, the overall conclusion then reads:

$$\chi\left(\alpha+1, j_{\min}\right) \underset{j_{\min}\to\infty}{\simeq} \frac{1}{p}\psi\left(\alpha+1, \overline{\imath}_{(\alpha^{ap})}(j_{\min})\right) \tag{A.14}$$

$$\underset{j_{\min}\to\infty}{\simeq} \frac{1}{p}\zeta\left(\alpha+1, \overline{\imath}_{(\alpha^{ap})}(j_{\min})\right). \tag{A.15}$$

A.3 EM approach for the estimation of α

Following the lines of [Duffield et al., 2003], we show in this section that the EM approach to the estimation of α from sampled data effectively leads to the same procedure as performing a fixed point method on likelihood solution of equation (6.22), as claimed in Section 6.3.2 of Chapter 6. The calculations are presented in the case $j_{\min} = 0$.

The EM algorithm was introduced in [Dempster et al., 1977], and a detailed desciption can be found in [McLachlan and Krishnan, 1997]. This versatile algorithm is an iterative procedure that converges to the maximum likelihood estimate under mild assumptions on the likelihood function (see [McLachlan and Krishnan, 1997]), when missing data prohibit its straightforward calculation. Observed data are said *incomplete* and viewed as an observable function of the so-called *complete data*. A proper choice of the *incomplete data* is a key to ensure tractability of the method and an adequate form of the iterative procedure. In our case, the incomplete data is simply the joint probability of S and Y, that we denote to simplify the expressions:

$$p_{ij} = P_{S,Y}(S = i \cap Y = j). \tag{A.16}$$

The observation η_j can easily be recovered from the incomplete data: $\eta_j = \sum_i p_{ij}$.

Assuming the complete to be known, we form the complete log-likelihood (recall that n is the number of observed flows):

$$\begin{aligned}\mathcal{L}_c(\alpha) &= \sum_{i=1}^{\infty}\sum_{j=0}^{\infty} p_{ij}\log\left(\phi_i B_p(i,j)\right) \\ &= n\log\zeta(\alpha+1) + n\sum_{i=1}^{\infty}\sum_{j=0}^{\infty} p_{ij}\log\left(B_p(i,j)i^{-(\alpha+1)}\right), \end{aligned} \tag{A.17}$$

where the sum over i, which should start at $i = j$ has been equivalently taken from 1 because $p_{ij} = 0$ if $i < j$.

After an initialization (*e.g.* $\alpha^{(0)} = 1$), the EM algorithm iterates two steps: the E-step (the expectation step) and the M-step (the maximization step).

E-step Form the expectation $\mathcal{Q}(\alpha, \alpha^{(k)})$ of the complete log likelihood, conditionally to the observation $\boldsymbol{\eta} = \{\eta_j, j = 1, 2, \cdots\}$, considering the parameter at the k-th iteration $\alpha^{(k)}$:

$$\mathcal{Q}(\alpha, \alpha^{(k)}) = \mathbb{E}_{\alpha^{(k)}}\{\mathcal{L}_c(\alpha)|\boldsymbol{\eta}\}. \tag{A.18}$$

M-step Define and determine:

$$\alpha^{(k+1)} = \underset{\alpha}{\operatorname{argmax}} \mathcal{Q}(\alpha, \alpha^{(k)}). \tag{A.19}$$

In our case, we have:

$$\mathcal{Q}(\alpha, \alpha^{(k)}) = n \log \zeta(\alpha + 1) + n \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \mathbb{E}_{\alpha^{(k)}}\{p_{ij}|\boldsymbol{\eta}\} \log \left(B_p(i, j) i^{-(\alpha+1)} \right). \tag{A.20}$$

The maximization step is obtain by simply solving the equation:

$$\frac{\mathrm{d}\mathcal{Q}(\alpha, \alpha^{(k)})}{\mathrm{d}\alpha} \mathcal{I}\alpha^{(k+1)} = 0, \tag{A.21}$$

where

$$\begin{aligned} \frac{\mathrm{d}\mathcal{Q}(\alpha, \alpha^{(k)})}{\mathrm{d}\alpha} &= n \frac{\zeta'(\alpha + 1)}{\zeta(\alpha + 1)} + n \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \mathbb{E}_{\alpha^{(k)}}\{p_{ij}|\boldsymbol{\eta}\} \ln i \\ &= n \frac{\zeta'(\alpha + 1)}{\zeta(\alpha + 1)} + n \sum_{i=1}^{\infty} \mathbb{E}_{\alpha^{(k)}}\{\phi_i|\boldsymbol{\eta}\} \ln i, \end{aligned} \tag{A.22}$$

noticing that $\sum_j p_{ij} = \phi_i$. Finally, direct computation yields:

$$\mathbb{E}_{\alpha^{(k)}}\{\phi_i|\boldsymbol{\eta}\} = \sum_{j=0}^{\infty} \eta_j \frac{\phi_i^{(k)} B_p(i, j)}{\sum_{l=1}^{\infty} \phi_l^{(k)} B_p(l, j)}, \tag{A.23}$$

so that plugging equation A.23 into equation A.22 brings (see the definition of $\bar{\imath}$ in equation 6.17 and recall that the summation on i in this definition can start indifferently at j or at 1 due to the presence of the term $B_p(i, j)$):

$$\begin{aligned} \frac{\mathrm{d}\mathcal{Q}(\alpha, \alpha^{(k)})}{\mathrm{d}\alpha} &= n \frac{\zeta'(\alpha + 1)}{\zeta(\alpha + 1)} + n \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \eta_j \frac{\phi_i^{(k)} B_p(i, j) \ln i}{\sum_{l=1}^{\infty} \phi_l^{(k)} B_p(l, j)} \\ &= n \frac{\zeta'(\alpha + 1)}{\zeta(\alpha + 1)} + n \sum_{j=0}^{\infty} \eta_j \ln \bar{\imath}_{(\alpha^{(k)})}(j). \end{aligned} \tag{A.24}$$

Resolution of equation A.21 then readily brings the equation:

$$\frac{\zeta'(\alpha^{(k+1)} + 1)}{\zeta(\alpha^{(k+1)} + 1)} = -\sum_{j=0}^{\infty} \eta_j \ln \bar{\imath}_{(\alpha^{(k)})}(j), \tag{A.25}$$

effectively is identical to the fix point resolution of equation (6.22) of Section 6.3.2, which concludes the demonstration.

Appendix B

Publications

International refereed journals

[1] Patrick Loiseau, Paulo Gonçalves, Guillaume Dewaele, Pierre Borgnat, Patrice Abry, and Pascale Vicat-Blanc Primet. Investigating self-similarity and heavy-tailed distributions on a large scale experimental facility. *IEEE/ACM Transactions on Networking*, December 2009. to appear.

[2] Edmundo Pereira de Souza Neto, Patrice Abry, Patrick Loiseau, Jean-Christophe Cejka, Marc-Antoine Custaud, Jean Frutoso, Claude Gharib, and Patrick Flandrin. Empirical mode decomposition to assess cardiovascular autonomic control in rats. *Fundamental & Clinical Pharmacology*, 21(5):481–496, October 2007.

International refereed conferences

[3] Patrick Loiseau, Paulo Gonçalves, Stéphane Girard, Florence Forbes, and Pascale Vicat-Blanc Primet. Maximum likelihood estimation of the flow size distribution tail index from sampled packet data. In *ACM Sigmetrics*, June 2009.

[4] Patrick Loiseau, Paulo Gonçalves, Romaric Guillier, Matthieu Imbert, Yuetsu Kodama, and Pascale Vicat-Blanc Primet. Metroflux: A high performance system for analyzing flow at very fine-grain. In *TridentCom*, April 2009.

[5] Patrick Loiseau, Paulo Gonçalves, and Pascale Vicat-Blanc Primet. A comparative study of different heavy tail index estimators of the flow size from sampled data. In *MetroGrid Workshop, GridNets*, New York, USA, October 2007. ACM Press.

Research reports

[6] Patrick Loiseau, Paulo Gonçalves, Guillaume Dewaele, Pierre Borgnat, Patrice Abry, and Pascale Vicat-Blanc Primet. Investigating self-similarity and heavy-tailed distributions on a large scale experimental facility. Technical Report 6472, INRIA, March 2008.

[7] Patrick Loiseau, Paulo Gonçalves, and Pascale Vicat-Blanc Primet. Impact of the correlation between flow rates and durations on the large-scale properties of aggregate network traffic. Technical Report 7100, INRIA, November 2009.

Miscellaneous

[8] Patrick Loiseau, Paulo Gonçalves, Guillaume Dewaele, Pierre Borgnat, Patrice Abry, and Pascale Vicat-Blanc Primet. Vérification du lien entre auto-similarité et distributions à queues

lourdes sur un dispositif grande échelle, June 2008. 9 ième Atelier en Evaluation de Performances, Aussois, France.

[9] Patrick Loiseau, Paulo Gonçalves, Romaric Guillier, Matthieu Imbert, Oana Goga, Yuetsu Kodama, and Pascale Vicat-Blanc Primet. *Metroflux*: a high performance system for very fine-grain flow analysis, April 2009. Grid'5000 Spring School.

[10] Patrick Loiseau, Paulo Gonçalves, Yuetsu Kodama, and Pascale Vicat-Blanc Primet. Metroflux: A fully operational high speed metrology platform, September 2008. Euro-NF workshop: New trends in modeling, quantitative methods and measurements, in cooperation with Net-coop, THOMSON Paris Research Labs, France.

[11] Patrick Loiseau, Paulo Gonçalves, and Pascale Vicat-Blanc Primet. How TCP can kill self-similarity, September 2008. Euro-NF workshop: Traffic Engineering and Dependability in the Network of the Future, VTT, Finland.

Demonstrations

[12] Patrick Loiseau, Romaric Guillier, Oana Goga, Matthieu Imbert, Paulo Gonçalves, and Pascale Vicat-Blanc Primet. Automated traffic measurements and analysis in Grid5000, June 2009. ACM Sigmetrics/Performance demonstration contest (**Best Demonstration Award**).

References

Online references

[13] Cisco. Netflow. see http://www.cisco.com/en/US/products/ps6601/~products_ios_protocol_group_home.html.

[14] http://www.endace.com/.

[15] Grid5000. http://www.grid5000.fr/.

[16] GtrcNet. http://www.gtrc.aist.go.jp/gnet.

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

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

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

[20] Ipsumdump. http://www.cs.ucla.edu/~kohler/ipsumdump/.

[21] LOBSTER project. http://www.ist-lobster.org/.

[22] The network simulator – ns-2. http://www.isi.edu/nsnam/ns/.

References

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

[Abry et al., 2009] Abry, P., Borgnat, P., Ricciato, F., Scherrer, A., and Veitch, D. (2009). Revisiting an old friend: On the observability of the relation between long range dependence and heavy tail. *Telecommunication Systems, Special Issue on Traffic Modeling, Its Computations and Applications*. To appear.

[Abry et al., 2000] Abry, P., Flandrin, P., Taqqu, M., and Veitch, D. (2000). Wavelets for the analysis, estimation and synthesis of scaling data. In Park, K. and Willinger, W., editors, *Self-Similar Network Traffic and Performance Evaluation*, chapter 2. John Wiley & Sons, Inc.

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

[Abry et al., 2002b] Abry, P., Gonçalves, P., and Lévy-Véhel, J. (2002b). *Lois d'échelles, fractales et ondelettes*. Hermes. English translation: Scaling, fractals and wavelets published by Wiley.

[Abry and Veitch, 1998] Abry, P. and Veitch, D. (1998). Wavelet analysis of long-range dependent traffic. *IEEE Transactions on Information Theory*, 44(1):2–15.

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

[Ait-hellal et al., 1997] Ait-hellal, O., Altman, E., Elouadghiri, D., Erramdani, M., and Mikou, N. (1997). Performance of TCP/IP: the case of two controlled sources. In *ICCC*.

[Altman et al., 2000] Altman, E., Avrachenkov, K., and Barakat, C. (2000). A stochastic model of TCP/IP with stationary random losses. In *ACM SIGCOMM*.

[Altman et al., 2005] Altman, E., Avrachenkov, K., and Barakat, C. (2005). A stochastic model of TCP/IP with stationary random losses. *IEEE/ACM Trans. Net.*, 13(2):356–369.

[Anderson and Arlitt, 2006] Anderson, E. and Arlitt, M. (2006). Full packet capture and offline analysis on 1 and 10 gb networks. Technical Report HPL-2006-156, HP.

[Arvidsson and Karlsson, 1999] Arvidsson, A. and Karlsson, P. (1999). On traffic models for tcp/ip. In *Teletraffic engineering in a competitive world, ITC-16*.

[Baccelli and Bonald, 1999] Baccelli, F. and Bonald, T. (1999). Window flow control in fifo networks with cross traffic. *Queueing Syst. Theory Appl.*, 32(1/3):195–231.

[Baccelli and Brémaud, 2003] Baccelli, F. and Brémaud, P. (2003). *Elements of Queueing Theory*. Springer.

[Baccelli and Hong, 2000] Baccelli, F. and Hong, D. (2000). Tcp is max-plus linear and what it tells us on its throughput. *SIGCOMM Comput. Commun. Rev.*, 30(4):219–230.

[Baccelli and Hong, 2002] Baccelli, F. and Hong, D. (2002). AIMD, fairness and fractal scaling of TCP traffic. In *IEEE INFOCOM*.

[Baccelli and Hong, 2005] Baccelli, F. and Hong, D. (2005). Interaction of tcp flows as billiards. *IEEE/ACM Trans. Netw.*, 13(4):841–853.

[Baccelli et al., 2007] Baccelli, F., Kim, K. B., and Mcdonald, D. R. (2007). Equilibria of a class of transport equations arising in congestion control. *Queueing Syst. Theory Appl.*, 55(1):1–8.

[Baccelli and McDonald, 2006] Baccelli, F. and McDonald, D. R. (2006). A stochastic model for the throughput of non-persistent tcp flows. In *valuetools '06: Proceedings of the 1st international conference on Performance evaluation methodolgies and tools*, page 58, New York, NY, USA. ACM.

[Barakat, 2001] Barakat, C. (2001). Tcp/ip modeling and validation. *IEEE Network*, 15(3):38–47.

[Barakat et al., 2005] Barakat, C., Iannaccone, G., and Diot, C. (2005). Ranking flows from sampled traffic. In *CoNEXT '05: Proceedings of the 2005 ACM conference on Emerging network experiment and technology*, pages 188–199, New York, NY, USA. ACM.

[Barakat et al., 2002] Barakat, C., Thiran, P., Iannaccone, G., Diot, C., and Owezarski, P. (2002). A flow-based model for internet backbone traffic. In *IMW '02: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, pages 35–47, New York, NY, USA. ACM.

[Barral and Gonçalves, 2009] Barral, J. and Gonçalves, P. (2009). Large deviations estimator for multiplicative cascades. Projet DMASC.

[Barral and Lévy Véhel, 2004] Barral, J. and Lévy Véhel, J. (2004). Multifractal analysis of a class of additive processes with correlated non-stationary increments. *Elec. J. Probab.*, 9:508–543.

[Barral and Mandelbrot, 2002] Barral, J. and Mandelbrot, B. (2002). Multifractal products of cylindrical pulses. *Probab. Theory Relat. Fields*, 124:409–430.

[Bavier et al., 2004] Bavier, A., Bowman, M., Chun, B., Culler, D., Karlin, S., Peterson, L., Roscoe, T., Spalink, T., and Wawrzoniak, M. (2004). Operating system support for planetary-scale network services. In *Proc. of the 1st Symposium on Network System Design and Implementation*.

[Baxter et al., 1991] Baxter, J., Jain, N., and R.S., V. (1991). Some familiar examples for which the large deviation principle does not hold. *Communications on pure and applied mathematics*, 44(8-9):911–923.

[Ben Azzouna et al., 2004] Ben Azzouna, N., Clerot, F., Fricker, C., and Guillemin, F. (2004). A flow-based approach to modeling adsl traffic on an ip backbone link. *Annals of Telecommunications*, 59(11/12):1260–1299.

[Ben Fredj et al., 2001] Ben Fredj, S., Bonald, T., Proutiere, A., Régnié, G., and Roberts, J. W. (2001). Statistical bandwidth sharing: a study of congestion at flow level. In *SIGCOMM '01: Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 111–122, New York, NY, USA. ACM.

[Benaïm and El Karoui, 2005] Benaïm, M. and El Karoui, N. (2005). *Promenade aléatoire: Chaines de Markov et simulation; martingales et stratégies*. Editions de l'école polytechnique.

[Beran, 1994] Beran, J. (1994). *Statistics for Long-memory processes*. Chapman & Hall, New York.

[Beran et al., 1995] Beran, J., Sherman, R., Taqqu, M. S., and Willinger, W. (1995). Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on communications*, 43(2-4):1566–1579.

[Bingham et al., 1987] Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1987). *Regular variations*. Cambridge University Press, Cambridge, UK.

[Blake et al., 1998] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and Weiss, W. (1998). An Architecture for Differentiated Service. RFC 2475 (Informational). Updated by RFC 3260.

[Blanc et al., 2009a] Blanc, A., Avrachenkov, K., and Collange, D. (2009a). Comparing some high speed TCP versions under bernoulli losses. In *PFLDnet*.

[Blanc et al., 2009b] Blanc, A., Avrachenkov, K., Collange, D., and Neglia, G. (2009b). Compound tcp with random losses. In *Networking, in LNCS*, volume 5550, pages 482–494.

[Bolze et al., 2006] Bolze, R., Cappello, F., Caron, E., Daydé, M., Desprez, F., Jeannot, E., Jégou, Y., Lanteri, S., Leduc, J., Melab, N., Mornet, G., Namyst, R., Primet, P., Quetier, B., Richard, O., Talbi, E.-G., and Touche, I. (2006). Grid'5000: a large scale and highly reconfigurable experimental grid testbed. *Int. J. of High Performance Computing Applications*, 20(4):481–494.

[Bonald, 1998] Bonald, T. (1998). Comparison of TCP reno and TCP vegas via fluid approximation. Technical Report 3563, INRIA.

[Bonald et al., 2002] Bonald, T., Oueslati-Boulahia, S., and Roberts, J. (2002). IP traffic and QoS control: the need for a flow-aware architecture. In *World Telecommunications Congress*.

[Borgnat et al., 2009] Borgnat, P., Dewaele, G., Fukuda, K., Abry, P., and Cho, K. (2009). Seven years and one day: Sketching the evolution of internet traffic. In *IEEE INFOCOM*.

[Bowen, 1975] Bowen, R. (1975). *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*. Springer.

[Braden et al., 1994] Braden, R., Clark, D., and Shenker, S. (1994). Integrated Services in the Internet Architecture: an Overview. RFC 1633 (Informational).

[Bradley, 2005] Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. 2:107–144.

[Brauckhoff et al., 2006] Brauckhoff, D., Tellenbach, B., Wagner, A., May, M., and Lakhina, A. (2006). Impact of packet sampling on anomaly detection metrics. In *IMC '06: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 159–164, New York, NY, USA. ACM.

[Brichet et al., 1996] Brichet, F., Roberts, J., Simonian, A., and Veitch, D. (1996). Heavy traffic analysis of a storage model with long range dependent on/off sources. *Queueing Systems*, 23:197–225.

[Bryc and Dembo, 1996] Bryc, W. and Dembo, A. (1996). Large deviations and strong mixing. *Ann. Inst. H. Poincaré Probab. Stat.*, 32:549–569.

[Budhiraja et al., 2004] Budhiraja, A., Hernández-Campos, F., Kulkarni, V. G., and Smith, F. D. (2004). Stochastic differential equation for tcp window size: Analysis and experimental validation. *Probab. Eng. Inf. Sci.*, 18(1):111–140.

[Cao et al., 2001a] Cao, J., Cleveland, W. S., Lin, D., and Sun, D. X. (2001a). The effect of statistical multiplexing on the long range dependence of internet packet traffic. Technical report, Bell Labs.

[Cao et al., 2001b] Cao, J., Cleveland, W. S., Lin, D., and Sun, D. X. (2001b). Internet traffic tends to poisson and independent as the load increases. Technical report, Bell Labs.

[Carofiglio et al., 2007] Carofiglio, G., Garetto, M., Leonardi, E., Tarello, A., and Marsan, M. A. (2007). Beyond fluid models: modelling tcp mice in ip networks under non-stationary random traffic. *Comput. Netw.*, 51(1):114–133.

[Casetti and Meo, 2000] Casetti, C. and Meo, M. (2000). A new approach to model the stationary behavior of TCP connections. In *IEEE INFOCOM*, pages 367–375.

[Chabchoub, 2009] Chabchoub, Y. (2009). *Analyse et modélisation du trafic Internet*. PhD thesis, Université Pierre et Marie Curie - Paris VI.

[Chabchoub et al., 2007] Chabchoub, Y., Fricker, C., Guillemin, F., and Robert, P. (2007). Deterministic versus probabilistic packet sampling in the internet. In *ITC'20*.

[Chabchoub et al., 2008] Chabchoub, Y., Fricker, C., Guillemin, F., and Robert, P. (2008). Inference of flow statistics via packet sampling in the internet. *IEEE Communications Letters*, 12(12):897–899.

[Chabchoub et al., 2009] Chabchoub, Y., Fricker, C., Guillemin, F., and Robert, P. (2009). On the statistical characterization of flows in internet traffic with application to sampling. *Computer Communications*. To Appear, see arXiv:0902.1736v2.

[Chainais et al., 2005] Chainais, P., Riedi, R., and Abry, P. (2005). On non-scale-invariant infinitely divisible cascades. *IEEE transactions on information theory*, 51(3):1063–1083.

[Clauset et al., 2009] Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*. to appear, see arXiv:0706.1062v2.

[Cox, 1984] Cox, D. (1984). Long range dependence: A review. In *Statistics: An Appraisal*, pages 55–74, Ames, IA, USA. Iowa State University Press.

[Cox and Isham, 1980] Cox, D. R. and Isham, V. (1980). *Point Processes*. Chapman & Hall.

[Crovella and Bestavros, 1996] Crovella, M. E. and Bestavros, A. (1996). Self-similarity in World Wide Web traffic: Evidence and possible causes. In *ACM SIGMETRICS*, pages 160–169.

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

[Crovella and Taqqu, 1999] Crovella, M. E. and Taqqu, M. S. (1999). Estimating the heavy tail index from scaling properties. *Methodology and Computing in Applied Probability*, 1(1):55–79.

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

[Daley and Vere-Jones, 2003] Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods*. Springer, 2nd edition.

[Daley and Vere-Jones, 2007] Daley, D. J. and Vere-Jones, D. (2007). *An Introduction to the Theory of Point Processes, Volume II: General Theory and Structure*. Springer, 2nd edition.

[Daubechies, 1992] Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM.

[D'Auria and Resnick, 2006] D'Auria, B. and Resnick, S. I. (2006). Data network models of burstiness. *Advances in Applied Probabilty*, 38(2):373–404.

[D'Auria and Resnick, 2008] D'Auria, B. and Resnick, S. I. (2008). The influence of dependence on data network models. *Advances in Applied Probabilty*, 40(1):60–94.

[Dembo and Zeitouni, 1998] Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*. Springer.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the *em* algorithm. *J. Roy. Stat. Soc., Series B (Method.)*.

[Dewaele et al., 2007] Dewaele, G., Borgnat, P., and Abry, P. (2007). Iptools : Analyse de trafic par sketch multi-résolution. logiciel déposé à l'APP (Association de Protection des Programmes) par le CNRS & ENS de Lyon, juillet 2007 (IDDN.FR.001.330007.000.S.P.2007.000.20700).

[Doukhan, 1994] Doukhan, P. (1994). *Mixing: Properties and Examples*. Springer.

[Doukhan et al., 2003] Doukhan, P., Oppenheim, G., and Taqqu, M. (2003). *Long-Range Dependence: Theory and Applications*. Birkhäuser, Boston.

[Duffield et al., 2003] Duffield, N., Lund, C., and Thorup, M. (2003). Estimating flow distributions from sampled flow statistics. In *SIGCOMM*.

[Dukkipati and McKeown, 2006] Dukkipati, N. and McKeown, N. (2006). Why flow-completion time is the right metric for congestion control. *SIGCOMM Comput. Commun. Rev.*, 36(1):59–62.

[Efron and Tibshirani, 1993] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

[Ellis, 1984] Ellis, R. S. (1984). Large deviations for a general class of random vectors. *Annals of Probability*, 12(1):1–12.

[Embrechts and Maejima, 2002] Embrechts, P. and Maejima, M. (2002). *Selfsimilar Processes*. Princeton University Press.

[Erramilli et al., 2000] Erramilli, A., Narayan, O., Neidhardt, A., and Saniee, I. (2000). Performance impacts of multi-scaling in wide area tcp/ip traffic. In *Proceedings of IEEE INFOCOM*, pages 352–359.

[Erramilli et al., 1996] Erramilli, A., Narayan, O., and Willinger, W. (1996). Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Trans. Netw.*, 4(2):209–223.

[Erramilli et al., 2002] Erramilli, A., Roughan, M., Veitch, D., and Willinger, W. (2002). Self-similar traffic and network dynamics. In *Proceedings of the IEEE*, pages 800–819.

[Estan and Varghese, 2002] Estan, C. and Varghese, G. (2002). New directions in traffic measurement and accounting. In *SIGCOMM*.

[Fan and Feng, 2000] Fan, A.-H. and Feng, D.-J. (2000). On the distribution of long-term time averages on symbolic space. *J. of Stat. Phys.*, 99(3-4):813–856.

[Feldmann et al., 1999a] Feldmann, A., Gilbert, A. C., Huang, P., and Willinger, W. (1999a). Dynamics of IP traffic: A study of the role of variability and the impact of control. In *ACM SIGCOMM*, pages 301–313.

[Feldmann et al., 1998a] Feldmann, A., Gilbert, A. C., and Willinger, W. (1998a). Data networks as cascades: Investigating the multifractal nature of internet WAN traffic. In *ACM SIGCOMM*, pages 42–55.

[Feldmann et al., 1999b] Feldmann, A., Gilbert, A. C., and Willinger, W. (1999b). Scaling analysis of conservative cascades, with applications to network traffic. *IEEE Trans. Info. Theo.*, 45(3):971–991.

[Feldmann et al., 1998b] Feldmann, A., Gilbert, A. C., Willinger, W., and Kurtz, T. (1998b). The changing nature of network traffic: Scaling phenomena. *Computer Communication Review*, 28:5–29.

[Feller, 1971] Feller, W. (1971). *An introduction to probability theory and its applications*, volume II. John Wiley & Sons, third edition.

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

[Figueiredo et al., 2002] Figueiredo, D. R., Liu, B., Misra, V., and Towsley, D. (2002). On the autocorrelation structure of TCP traffic. *Computer Networks*, 40(3):339–361.

[Fiorini et al., 1997] Fiorini, P., Lipsky, L., and Crovella, M. (1997). Consequences of ignoring self-similar data traffic in communications modeling. In *Proceedings of Tenth International Conference on Parallel and Distributed Computing Systems (PDCS-97)*, pages 322–327.

[Floyd, 2003] Floyd, S. (2003). RFC 3649: HighSpeed TCP for Large Congestion Windows. RFC 3649.

[Fortin-Parisi and Sericola, 2004] Fortin-Parisi, S. and Sericola, B. (2004). A markov model of TCP throughput, goodput and slow start. *Perf. Eval.*, 58(2+3):89–108.

[Gaigalas and Kaj, 2003] Gaigalas, R. and Kaj, I. (2003). Convergence of scaled renewal processes and a packet arrival model. *Bernoulli*, 9(4):671–703.

[Garetto and Towsley, 2003] Garetto, M. and Towsley, D. (2003). Modeling, simulation and measurements of queuing delay under long-tail internet traffic. In *ACM SIGMETRICS*, pages 47–57, New York, NY, USA. ACM.

[Garetto and Towsley, 2008] Garetto, M. and Towsley, D. (2008). An efficient technique to analyze the impact of bursty tcp traffic in wide-area networks. *Perform. Eval.*, 65(2):181–202.

[Garrett and Willinger, 1994] Garrett, M. W. and Willinger, W. (1994). Analysis, modeling and generation of self-similar vbr video traffic. *SIGCOMM Comput. Commun. Rev.*, 24(4):269–280.

[Gibbens et al., 2000] Gibbens, R. J., Sargood, S. K., Eijl, C. V., Kelly, F. P., Azmoodeh, H., Macfadyen, N. W., and Macfadyen, N. W. (2000). Fixed-point models for the end-to-end performance analysis of ip networks. In *in Proceedings of the 13th ITC Specialist Seminar: IP Traffic Measurement, Modeling and Management, Sept 2000.*

[Gonçalves and Riedi, 2005] Gonçalves, P. and Riedi, R. (2005). Diverging moments and parameter estimation. *J. American Stat. Assoc.*, 100(472):1382–1393.

[Gong et al., 2005] Gong, W.-B., Liu, Y., Misra, V., and Towsley, D. (2005). Self-similarity and long range dependence on the internet: a second look at the evidence, origins and implications. *Comput. Netw.*, 48(3):377–399.

[Greiner et al., 1999] Greiner, M., Jobmann, M., and Lipsky, L. (1999). The importance of power-tail distributions for modeling queueing systems. *Oper. Res.*, 47(2):313–326.

[Grossglauser and Bolot, 1999] Grossglauser, M. and Bolot, J.-C. (1999). On the relevance of long-range dependence in network traffic. *IEEE/ACM Trans. Netw.*, 7(5):629–640.

[Guerin et al., 2003] Guerin, C. A., Nyberg, H., Perrin, O., Resnick, S., Rootzén, H., and Starica, C. (2003). Empirical testing of the infinite source poisson data traffic model. *Stochastic Models*, 19(2):151–200.

[Guillier, 2009] Guillier, R. (2009). *Methodologies and Tools for the Evaluation of Transport Protocols in the Context of Highspeed Networks*. PhD thesis, ENS-Lyon, Université de Lyon.

[Guo et al., 2001] Guo, L., Crovella, M., and Matta, I. (2001). How does TCP generate pseudo-self-similarity? In *MASCOTS*, page 215, Washington, DC, USA. IEEE Computer Society.

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

[Heyman and Lakshman, 1996] Heyman, D. P. and Lakshman, T. V. (1996). What are the implications of long-range dependence for vbr-video traffic engineering? *IEEE/ACM Trans. Netw.*, 4(3):301–317.

[Heyman et al., 1997] Heyman, D. P., Lakshman, T. V., and Neidhardt, A. L. (1997). A new method for analysing feedback-based protocols with applications to engineering web traffic over the internet. *SIGMETRICS Perform. Eval. Rev.*, 25(1):24–38.

[Hill, 1975] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

[Hohn and Veitch, 2006] Hohn, N. and Veitch, D. (2006). Inverting sampled traffic. *IEEE/ACM Trans. Netw.*, 14(1):68–80.

[Hohn et al., 2002a] Hohn, N., Veitch, D., and Abry, P. (2002a). Does fractal scaling at the ip level depend on tcp flow arrival processes? In *IMW '02: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, pages 63–68, New York, NY, USA. ACM.

[Hohn et al., 2002b] Hohn, N., Veitch, D., and Abry, P. (2002b). Investigating the scaling behaviour of internet flow arrivals. In *Self-similarity and applications*.

[Hohn et al., 2003] Hohn, N., Veitch, D., and Abry, P. (2003). Cluster processes, a natural language for network traffic. *IEEE Transactions on Signal Processing – Special Issue on Signal Processing in Networking*, 8(51):2229–2244.

[Horváth and Telek, 2002] Horváth, A. and Telek, M. (2002). Markovian modeling of real data traffic: Heuristic phase type and map fitting of heavy tailed and fractal like samples. In *Performance Evaluation of Complex Systems: Techniques and Tools, Performance 2002, Tutorial Lectures*, pages 405–434, London, UK. Springer-Verlag.

[Hurley et al., 1999] Hurley, P., Le Boudec, J.-Y., and Thiran, P. (1999). A note on the fairness of additive increase and multiplicative decrease. In *ITC-16*.

[Jacobson, 1988] Jacobson, V. (1988). Congestion avoidance and control. In *SIGCOMM '88: Symposium proceedings on Communications architectures and protocols*, pages 314–329, New York, NY, USA. ACM.

[Jacobson et al., 1992] Jacobson, V., Braden, R., and Borman, D. (1992). TCP Extensions for High Performance. RFC 1323.

[Jaffard, 1999] Jaffard, S. (1999). The multifractal nature of lévy processes. *Probability theory and related fields*, 114(2):207–227.

[Jelenković and Momčilović, 2003] Jelenković, P. and Momčilović, P. (2003). Asymptotic loss probability in a finite buffer fluid queue with hetergeneous heavy-tailed on–off processes. *Ann. Appl. Probab.*, 13(2):576–603.

[Jelenković and Lazar, 1999] Jelenković, P. R. and Lazar, A. A. (1999). Asymptotic results for multiplexing subexponential on-off processes. *Adv. in Appl. Probab.*, 31(2):394–421.

[Jiang and Dovrolis, 2004a] Jiang, H. and Dovrolis, C. (2004a). The effect of flow capacities on the burstiness of aggregated traffic. In *Lecture Notes in Computer Science*, volume 3015, pages 93–102. Springer.

[Jiang and Dovrolis, 2004b] Jiang, H. and Dovrolis, C. (2004b). The origin of tcp traffic burstiness in some time scales. Technical report, in IEEE INFOCOM.

[Jiang and Dovrolis, 2005] Jiang, H. and Dovrolis, C. (2005). Why is the internet traffic bursty in short time scales? *SIGMETRICS Perform. Eval. Rev.*, 33(1):241–252.

[Joo et al., 2001] Joo, Y., Ribeiro, V., Feldmann, A., Gilbert, A. C., and Willinger, W. (2001). Tcp/ip traffic dynamics and network performance: a lesson in workload modeling, flow control, and trace-driven simulations. *SIGCOMM Comput. Commun. Rev.*, 31(2):25–37.

[Kaj and Olsén, 2001] Kaj, I. and Olsén, J. (2001). Throughput modeling and simulation for single connection tcp-tahoe. In *17th International Teletraffic Congress ITC'01*.

[Kamal, 2004] Kamal, A. E. (2004). Discrete-time modeling of TCP reno under background traffic interference with extension to red-based routers. *Perf. Eval.*, 58(2+3):109–142.

[Karagiannis et al., 2004a] Karagiannis, T., Faloutsos, M. M. M., and Broido, A. (2004a). A non stationary Poisson view of the internet traffic. In *INFOCOM*.

[Karagiannis et al., 2004b] Karagiannis, T., Molle, M., and Faloutsos, M. (2004b). Long-range dependence - ten years of internet traffic modeling. *IEEE Internet Computing*.

[Kawahara et al., 2007] Kawahara, R., Mori, T., Kamiyama, N., Harada, S., and Asano, S. (2007). A study on detecting network anomalies using sampled flow statistics. In *SAINT-W '07: Proceedings of the 2007 International Symposium on Applications and the Internet Workshops*, page 81, Washington, DC, USA. IEEE Computer Society.

[Kelly, 1999] Kelly, F. (1999). Mathematical modelling of the internet. In *Proc. of 4th Int. Congress on Industrial and Applied Mathematics*.

[Kelly et al., 1998] Kelly, F. P., Maulloo, A., and Tan, D. (1998). Rate control for communication networks: Shadow prices, proportional fairness and stability. *Journal of Operation Research*, 49(3):237–252.

[Khayari et al., 2004] Khayari, R. E. A., Sadre, R., Haverkort, B. R., and Ost, A. (2004). The pseudo-self-similar traffic model: application and validation. *Perform. Eval.*, 56(1-4):3–22.

[Kilpi and Norros, 2002] Kilpi, J. and Norros, I. (2002). Testing the gaussian approximation of aggregate traffic. In *IMW '02: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, pages 49–61, New York, NY, USA. ACM.

[Kingman, 1993] Kingman, J. (1993). *Poisson processes*. Oxford university press.

[Kleinrock, 1969] Kleinrock, L. (1969). Models for computer networks. In *Proceedings of the Int. Conf. on Communication*, Boulder, Colo.

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

[Kumar, 1998] Kumar, A. (1998). Comparative performance analysis of versions of TCP in a local network with a lossy link. *IEEE/ACM Trans. Net.*, 6(4):485–498.

[Kurtz, 1996] Kurtz, T. G. (1996). Limit theorems for workload input models. In *Stochastic networks: theory and applications*. Clarendon Press, Oxford.

[Lakshman and Madhow, 1997] Lakshman, T. V. and Madhow, U. (1997). The performance of tcp/ip for networks with high bandwidth-delay products and random loss. *IEEE/ACM Trans. Netw.*, 5(3):336–350.

[Le Boudec and Thiran, 2001] Le Boudec, J.-Y. and Thiran, P. (2001). *Network Calculus*. Lecture Notes in Computer Science (LNCS). Springer Verlag.

[Leland et al., 1993] Leland, W. E., Taqqu, M. S., Willinger, W., and Wilson, D. V. (1993). On the self-similar nature of ethernet traffic. In *ACM SIGCOMM*, pages 183–193, New York, NY, USA. ACM Press.

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

[Levy and Taqqu, 1987] Levy, J. B. and Taqqu, M. S. (1987). On renewal processes having stable inter-renewal intervals and stable rewards. *Les Annales des Sciences Mathématiques du Québec*, 11:95–110.

[Levy and Taqqu, 2000] Levy, J. B. and Taqqu, M. S. (2000). Renewal reward processes with heavy-tailed inter-renewal times and heavy-tailed rewards. *Bernoulli*, 6(1):23–44.

[Liu et al., 2006] Liu, W., Gong, J., Ding, W., and Cheng, G. (2006). A method for estimation of flow length distributions from sampled flow statistics. In *ICOIN*.

[Lowen and Teich, 1993] Lowen, S. B. and Teich, M. C. (1993). Fractal renewal processes generate 1/f noise. *Phys. Rev. E*, 47(2):992–1001.

[Lowen and Teich, 2005] Lowen, S. B. and Teich, M. C. (2005). *Fractal-Based Point Processes*. Wiley.

[MacIntyre et al., 2008] MacIntyre, S., Behnoodi, G., Marcondes, C., Gerla, M., and Cavendish, D. (2008). Assessing tcp protocols variants and file size impact on aggregate internet traffic statistics. In *PFLDnet*.

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

[Mandelbrot, 1969] Mandelbrot, B. (1969). Long-run linearity, locally gaussian process, h-spectra and infinite variances. *International Economic Review*, 10(1):82–111.

[Mandelbrot, 1997] Mandelbrot, B. (1997). *Fractals and Scaling In Finance*. Springer.

[Mandelbrot and Van Ness, 1968] Mandelbrot, B. and Van Ness, J. (1968). Fractional brownian motions, fractional noises and applications. *SIAM Rev.*, 10(4):422–437.

[Mandjes and Boots, 2001] Mandjes, M. and Boots, N. K. (2001). The shape of the loss curve and the impact of long-range dependence on network performance. Tinbergen Institute Discussion Papers 01-051/4, Tinbergen Institute.

[Mandjes and Kim, 2001] Mandjes, M. and Kim, J. H. (2001). Large deviations for small buffers: An insensitivity result. *Queueing Syst. Theory Appl.*, 37(4):349–362.

[Mandjes and Borst, 2000] Mandjes, M. R. and Borst, S. C. (2000). Overflow behavior in queues with many long-tailed inputs. *Advances in Applied Probability*, 32(4):1150–1167.

[Markovich, 2007] Markovich, N. (2007). *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*. Wiley.

[Massoulié and Roberts, 2000] Massoulié, L. and Roberts, J. (2000). Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems*, 15:185–201.

[Massoulié and Simonian, 1999] Massoulié, L. and Simonian, A. (1999). Large buffer asymptotics for the queue with fbm input. *Journal of Applied Probability*, 36:894–906.

[Mathis et al., 1997] Mathis, M., Semke, J., Mahdavi, J., and Ott, T. (1997). The macroscopic behavior of the TCP congestion avoidance algorithm. *SIGCOMM Comput. Commun. Rev.*, 27(3):67–82.

[Maulik and Resnick, 2003a] Maulik, K. and Resnick, S. (2003a). The self-similar and multifractal nature of a network traffic model. *Stochastic Models*, 19(4):549–577.

[Maulik and Resnick, 2003b] Maulik, K. and Resnick, S. (2003b). Small and large time scale analysis of a network traffic model. *Queueing Syst. Theory Appl.*, 43(3):221–250.

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

[McLachlan and Krishnan, 1997] McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley's Series in Probability and Statistics. Wiley, New York.

[Meyer, 1992] Meyer, Y. (1992). *Wavelets and operators*. Cambridge University Press.

[Mikosch et al., 2002] Mikosch, T., Resnick, S., Rootzén, H., and Stegeman, A. (2002). Is network traffic approximated by stable lévy motion or fractional brownian motion? *The Annals of Applied Probability*, 12(1):23–68.

[Mills, 1992] Mills, D. L. (1992). Network time protocol (version 3) specification, implementation and analysis. RFC 1035.

[Misra et al., 1999] Misra, V., Gong, W.-B., and Towsley, D. (1999). Stochastic differential equation modeling and analysis of TCP-windowsize behavior. In *Performance'99*.

[Misra et al., 1998] Misra, V., Misra, V., and bo Gong, W. (1998). A hierarchical model for teletraffic. In *Proceedings of the 37th Annual IEEE CDC*, pages 1674–1679.

[Mori et al., 2004] Mori, T., Uchida, M., Kawahara, R., Pan, J., and Goto, S. (2004). Identifying elephant flows through periodically sampled packets. In *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 115–120, New York, NY, USA. ACM.

[Narayan, 1998] Narayan, O. (1998). Exact asymptotic queue length distribution for fractional brownian traffic. In *Advances in Performance Analysis*, pages 39–63.

[Neidhardt and Wang, 1998] Neidhardt, A. L. and Wang, J. L. (1998). The concept of relevant time scales and its application to queuing analysis of self-similar traffic (or is hurst naughty or nice?). *SIGMETRICS Perform. Eval. Rev.*, 26(1):222–232.

[Nguyen et al., 2004] Nguyen, H., Thiran, P., and Barakat, C. (2004). On the correlation of tcp traffic in backbone networks. In *ISCAS (IEEE International Symposium on Circuits and Systems)*.

[Nolan, 2001] Nolan, J. P. (2001). Maximum likelihood estimation and diagnostics for stable distributions. In Barndorff-Nielsen, O., Mikosh, T., and Resnick, S., editors, *Lévy Processes : Theory and application*. Birkhäuser, Boston.

[Norros, 1994] Norros, I. (1994). A storage model with self-similar input. *Queueing Systems*, 16(3-4):387–396.

[Padhye et al., 1998] Padhye, J., Firoiu, V., Towsley, D., and Kurose, J. (1998). Modeling TCP throughput: a simple model and its empirical validation. *ACM SIGCOMM CCR*, 28(4):303–314.

[Padhye et al., 2000] Padhye, J., Firoiu, V., Towsley, D. F., and Kurose, J. F. (2000). Modeling TCP reno performance: a simple model and its empirical validation. *IEEE/ACM Trans. Net.*, 8(2):133–145.

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

[Park et al., 1997] Park, K., Kim, G., and Crovella, M. (1997). On the effect of traffic self-similarity on network performances. In *SPIE International Conference on Performance and Control of Network Systems*.

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

[Paxson and Allman, 2000] Paxson, V. and Allman, M. (2000). Computing TCP's retransmission timer. RFC 2988.

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

[Paxson and Floyd, 1995] Paxson, V. and Floyd, S. (1995). Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3:226–244.

[Polychronakis et al., 2004] Polychronakis, M., Anagnostakis, K. G., Markatos, E. P., and ysleb, A. (2004). Design of an application programming interface for ip network monitoring. In *Proc. of the 9th IFIP/IEEE Network Operations and Management Symposium (NOMS04)*, pages 483–496.

[Postel, 1981] Postel, J. (1981). Transmission Control Protocol. RFC 793.

[Prasad and Dovrolis, 2008] Prasad, R. S. and Dovrolis, C. (2008). Beyond the model of persistent tcp flows: Open-loop vs closed-loop arrivals of non-persistent flows. In *ANSS-41 '08: Proceedings of the 41st Annual Simulation Symposium (anss-41 2008)*, pages 121–130, Washington, DC, USA. IEEE Computer Society.

[Qiu et al., 2001] Qiu, L., Zhang, Y., and Keshav, S. (2001). Understanding the performance of many TCP flows. *Comp. Net.*, 37(3-4):277–306.

[Revuz, 1984] Revuz, D. (1984). *Markov chains*. Elseiver.

[Rhee and Xu, 2005] Rhee, I. and Xu, L. (2005). CUBIC: A New TCP-Friendly High-Speed TCP Variants. In *PFLDnet*.

[Ribeiro et al., 2006a] Ribeiro, B., Towsley, D., Ye, T., and Bolot, J. C. (2006a). Fisher information of sampled packets: an application to flow size estimation. In *IMC '06: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 15–26, New York, NY, USA. ACM.

[Ribeiro et al., 2006b] Ribeiro, V. J., Riedi, R. H., and Baraniuk, R. G. (2006b). Multiscale queueing analysis. *IEEE/ACM Trans. Netw.*, 14(5):1005–1018.

[Ribeiro et al., 2005] Ribeiro, V. J., Zhang, Z.-L., Moon, S., and Diot, C. (2005). Small-time scaling behavior of internet backbone traffic. *Comp. Net.*, 48(3):315–334.

[Riedi, 2003] Riedi, R. H. (2003). Multifractal processes. In Doukhan, P., Oppenheim, G., and Taqqu, M. S., editors, *Theory and Applications of Long-Range Dependence*, pages 223–233. Birkhäuser, Boston, MA, USA.

[Riedi et al., 1999] Riedi, R. H., Crouse, M. S., Ribeiro, V. J., and Baraniuk, R. G. (1999). A multifractal wavelet model with application to network traffic. *IEEE Trans. Info. Theo.*, 45(3):992–1018.

[Riedi and Lévy Véhel, 1997] Riedi, R. H. and Lévy Véhel, J. (1997). Multifractal properties of TCP traffic: a numerical study. Technical report, INRIA.

[Rio, 2000] Rio, E. (2000). *Théorie asymptotique des processus aléatoires faiblement dépendants*. Springer.

[Robert and Le Boudec, 1997] Robert, S. and Le Boudec, J.-Y. (1997). New models for pseudo self-similar traffic. *Perform. Eval.*, 30(1-2):57–68.

[Roberts, 2004] Roberts, J. W. (2004). A survey on statistical bandwidth sharing. *Comput. Netw.*, 45(3):319–332.

[Rockafellar, 1970] Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

[Rodolakis and Jacquet, 2005] Rodolakis, G. and Jacquet, P. (2005). An analytical evaluation of autocorrelations in TCP traffic. In Cho, K. and Jacquet, P., editors, *Technologies for Advanced Heterogeneous Networks*, volume 3837 of *Lecture Notes in Computer Science*, pages 296–306. Springer.

[Roughan et al., 2001] Roughan, M., Erramilli, A., and Veitch, D. (2001). Network performance for tcp networks, part i: Persistent sources. In *Proc. 17th Int. Teletraffic Congress*.

[Roughan and Gottlieb, 2002] Roughan, M. and Gottlieb, J. (2002). Large-scale measurement and modeling of backbone internet traffic. In *SPIE ITCom*.

[Roughan and Veitch, 2007] Roughan, M. and Veitch, D. (2007). Some remarks on unexpected scaling exponents. *CCR Online*.

[Roughan et al., 1999] Roughan, M., Yates, J., and Veitch, D. (1999). 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.

[Ruelle, 1984] Ruelle, D. (1984). *Thermodynamic formalism. The mathematical structures of classical equilibrium statistical mechanics*. Cambridge University Press.

[Ryu and Lowen, 1996] Ryu, B. and Lowen, S. (1996). Point process approaches to the modeling and analysis ofself-similar traffic – part i. model construction. In *IEEE INFOCOM*.

[Ryu and Elwalid, 1996] Ryu, B. K. and Elwalid, A. (1996). The importance of long-range dependence of vbr video traffic in atm traffic engineering: myths and realities. *SIGCOMM Comput. Commun. Rev.*, 26(4):3–14.

[Samorodnitsky and Taqqu, 1994] Samorodnitsky, G. and Taqqu, M. S. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman & Hall.

[Sarvotham et al., 2001] Sarvotham, S., Riedi, R., and Baraniuk, R. (2001). Connection-level analysis and modeling of network traffic. In *IMW '01: Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, pages 99–103, New York, NY, USA. ACM.

[Sarvotham et al., 2005] Sarvotham, S., Riedi, R., and Baraniuk, R. (2005). Network and user driven alpha-beta on-off source model for network traffic. *Comput. Netw.*, 48(3):335–350.

[Scherrer et al., 2007] Scherrer, A., Larrieu, N., Owezarski, P., Borgnat, P., and Abry, P. (2007). Non-gaussian and long memory statistical characterizations for internet traffic with anomalies. *IEEE Trans. Dependable Secur. Comput.*, 4(1):56–70.

[Schneider et al., 2007] Schneider, F., Wallerich, J., and Feldmann, A. (2007). Packet Capture in 10-Gigabit Ethernet Environments Using Contemporary Commodity Hardware. In *Passive and Active Network Measurement: 8th International Conference, Pam 2007, Louvain-la-neuve, Belgium, April 5-6, 2007, Proceedings*. Springer.

[Schroeder et al., 2006] Schroeder, B., Wierman, A., and Harchol-Balter, M. (2006). Open versus closed: a cautionary tale. In *NSDI'06: Proceedings of the 3rd conference on Networked Systems Design & Implementation*, pages 18–18, Berkeley, CA, USA. USENIX Association.

[Seal, 1952] Seal, H. L. (1952). The maximum likelihood fitting of the discrete Pareto law. *Journal of the Institute of Actuaries*, 78:115–121.

[Sikdar et al., 2003] Sikdar, B., Kalyanaraman, S., and Vastola, K. S. (2003). Analytic models for the latency and steady-state throughput of TCP tahoe, reno, and sack. *IEEE/ACM Trans. Net.*, 11(6):959–971.

[Sikdar and Vastola, 2001] Sikdar, B. and Vastola, K. S. (2001). The effect of tcp on the self-similarity of network traffic. In *In Proc. of the 35th Conf. on Information Sciences and Systems*, pages 21–23.

[Simonian and Veitch, 1997] Simonian, A. and Veitch, D. (1997). A storage model with high rate and long range dependent on/off sources sources. available at http://www.cubinlab.ee.unimelb.edu.au/ darryl/Publications/OnOffstorage.pdf.

[Soudan, 2009] Soudan, S. (2009). *Bandwidth Sharing and Control in High-Speed Networks: Combining Packet- and Circuit-Switching Paradigms*. PhD thesis, ENS-Lyon, Université de Lyon.

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

[Srikant, 2007] Srikant, R. (2007). Keynote speach, PFLDNet 2007.

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

[Taqqu et al., 1997a] Taqqu, M., Teverovsky, V., and Willinger, W. (1997a). Is network traffic self-similar or multifractal? *Fractals*, 5:63–73.

[Taqqu and Levy, 1986] Taqqu, M. S. and Levy, J. B. (1986). Using renewal processes to generate long-range dependence and high variability. In *Dependence in Probability and Statistics*, pages 73–89. Birkhäuser, Boston.

[Taqqu et al., 1997b] Taqqu, M. S., Willinger, W., and Sherman, R. (1997b). Proof of a fundamental result in self-similar traffic modeling. *ACM SIGCOMM Computer Communication Review*, 27(2):5–23.

[Tune and Veitch, 2008] Tune, P. and Veitch, D. (2008). Towards optimal sampling for flow size estimation. In *IMC '08: Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, pages 243–256, New York, NY, USA. ACM.

[Veitch et al., 2005] Veitch, D., Hohn, N., and Abry, P. (2005). Multifractality in TCP/IP traffic: the case against. *Comp. Net.*, 48(3):293–313.

[Veres and Boda, 2000] Veres, A. and Boda, M. (2000). The chaotic nature of tcp congestion control. In *IEEE INFOCOM*, pages 1715–1723.

[Veres et al., 2000] Veres, A., Kenesi, Z., Molnár, S., Vattay, G., and /d, P. P. (2000). On the propagation of long-range dependence in the internet. In *ACM SIGCOM*.

[Wang et al., 2002] Wang, X., Sarvotham, S., Riedi, R., and Baraniuk, R. (2002). Network traffic modeling using connection-level information. In *SPIE ITCom*.

[White et al., 2002] White, B., Lepreau, J., Stoller, L., Ricci, R., Guruprasad, S., Newbold, M., Hibler, M., Barb, C., and Joglekar, A. (2002). An integrated experimental environment for distributed systems and networks. *ACM SIGOPS Operating Systems Review*, 36(SI):255–270.

[Wierman et al., 2003a] Wierman, A., Osogami, T., and Olsén, J. (2003a). Modeling tcp-vegas under on/off traffic. *MAMA Workshop, SIGMETRICS Perform. Eval. Rev.*, 31(2):6–8.

[Wierman et al., 2003b] Wierman, A., Osogami, T., and Olsén, J. (2003b). A unified framework for modeling TCP-Vegas, TCP-SACK, and TCP-Reno. In *MASCOTS 2003*.

[Willinger et al., 1995] Willinger, W., Taqqu, M. S., Sherman, R., and Wilson, D. V. (1995). Self-similarity through high-variability: Statistical analysis of ethernet lan traffic at the source level. In *ACM SIGCOMM*, pages 100–113.

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

[Xu et al., 2004] Xu, L., Harfoush, K., and Rhee, I. (2004). Binary increase congestion control for fast long-distance networks. In *INFOCOM*.

[Yang and Michailidis, 2007] Yang, L. and Michailidis, G. (2007). Sampled based estimation of network traffic flow characteristics. In *INFOCOM*.

[Zhang et al., 2003] Zhang, Z.-L., Ribeiro, V. J., Moon, S., and Diot, C. (2003). Small-time scaling behavior of internet backbone traffic: an empirical study. In *IEEE INFOCOM*.

[Zinsmeister, 2000] Zinsmeister, M. (2000). *Thermodynamic formalism and holomorphic dynamical systems*. American Mathematical Society.