How TCP can kill Self-Similarity

Patrick Loiseau, Paulo Gonçalves, Pascale Primet Vicat-Blanc

I. Motivations

Comprehension of network traffic characteristics is a central preoccupation for Internet Service Providers. From a statistical viewpoint, this is a difficult problem since it encompasses several components as the network design and the transport protocol used (TCP, UDP).

Over the last decade, many research efforts have been devoted to the study of aggregated traffic time series collected at the core of networks. The pioneering works by [1], [2] showed that the Poisson hypothesis, which is relevantly used in phone networks, was not suitable to describe computer networks. Instead, self-similarity was proved a much more appropriate paradigm [1], [2], [3]. Then, the theoretical work from Taqqu and collaborators [4], [2] identified the heavy-tailed nature of the file size distribution as a possible origin for the observed self-similarity. In addition, it gave the exact relation between the self-similarity index and the tail index that should be observed when the sources behavior is modeled with the ON/OFF model. Despite a controversial debate on the question, it has then been more recently stated that the TCP congestion control mechanism cannot be responsible for the self-similarity observed in the large time scales [5], [6].

On the opposite side, we show in this work that when the file size is heavy-tailed, the TCP congestion control mechanism under sufficiently high loss can annihilate the self-similarity that would be observed without any loss.

II. Theory: brief overview

We consider the ON/OFF model with a large number of sources: each source "regularly" emits the packets of flows during the ON periods, which are separated by idle times (OFF periods).

The flow size W is said to be heavy-tailed, with tail exponent $\alpha > 0$ (and noted α-HT) when the tail of its cumulative distribution function, F_W, is characterized by an algebraic decrease [7]:

$$P(W > w) = 1 - F_W(w) \sim L(w) \cdot w^{-\alpha} \quad \text{for } w \to \infty, \quad (1)$$

where $L(w)$ is a slowly varying function (i.e. $\forall a > 0,\ L(aw)/L(w) \to_{w\to\infty} 1$). As a paradigm for α-HT positive random variable, we used the Pareto distribution:

$$F_W(w) = 1 - \left(\frac{k}{w+k}\right)^{\alpha}, \quad (2)$$

with $k > 0$ and $\alpha > 1$, which mean reads: $EW = k/(\alpha - 1)$.

Patrick Loiseau is with Université de Lyon, École Normale Supérieure de Lyon (LIP), France. (e-mail: Patrick.Loiseau@ens-lyon.fr)

Paulo Gonçalves and Pascale Vicat-Blanc Primet are with INRIA, Université de Lyon, École Normale Supérieure de Lyon (LIP), France. (e-mails: Paulo.Goncalves@ens-lyon.fr, Pascale.Primet@ens-lyon.fr)

The aggregated throughput $X(t)$ is said to be long-range dependent (LRD) if

$$EX(t)X(t+\tau) \underset{\tau\to\infty}{\sim} |\tau|^{2H-2}, \quad (3)$$

with $\frac{1}{2} < H < 1$.

Taqqu's theorem roughly states that in the ON/OFF model context with many sources, if the flow size distribution is α-HT (1) and the OFF time distribution is exponential, then the aggregated throughput will exhibit the LRD property in the large time scale with the index:

$$H = \frac{3 - \alpha^*}{2}, \quad (4)$$

where $\alpha^* = \min(\alpha, 2)$. The same results remains true if the OFF times are heavy-tailed distributed and the flow size distribution is exponential. If both distributions are heavy-tailed, the smaller tail index imposes the LRD index. If both distributions are exponential, there is no LRD ($H = 0.5$).

This theoretical prediction has been experimentally validated on computer networks simulators (ns2) [8], and on a real large scale computer network [6]. In the latter, the ideal case where there is no congestion (and consequently no packet loss) where considered. In this work, we tackle the question of the LRD observed under congestion using an experimental approach based on real large scale network experiments.

III. Experimental setup

We now present the experimental setup used for our experiments.

A. Grid5000

Grid5000 is a 5000 CPUs nation wide grid infrastructure dedicated to network and grid research [9]. It is constituted of 9 sites geographically distributed, which are interconnected with 10 Gbps dedicated optical network provided by RENATER, the French National Research and Education Network (see Figure 1). Each site is hosting one or more cluster of about 500 cores.

Grid5000 is a research tool, featured with high control, reconfiguration and monitoring capabilities. Thanks to the use of the reservation tool OAR, it allows the users to reserve the same set of nodes across successive experiments. The use of the environment deployment system Kadeploy then allows the users to automatically deploy their own customize environment on the reserved nodes. For example kernel modules for congestion control variants or QoS measurement can be added to the native operative system.

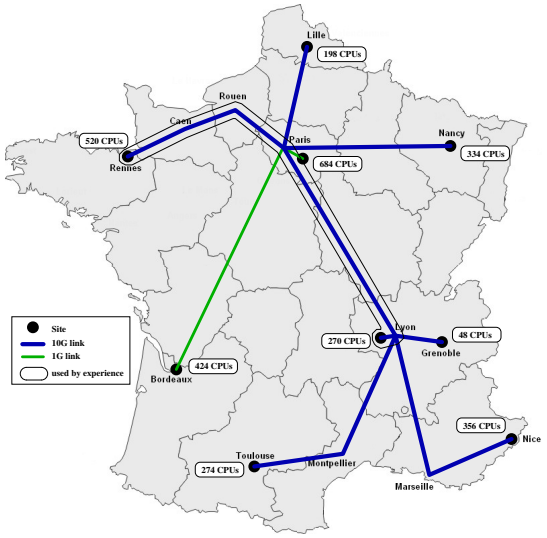

Fig. 1. Grid500

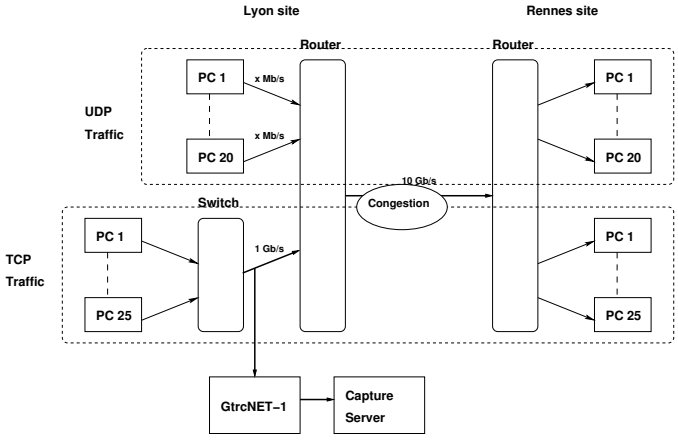

Fig. 2. Experimental topology

B. Experimental system description

The topology used for our experiments is described on Figure 2: 25 nodes in Lyon are emitting TCP flows to 25 nodes in Rennes. Each of these nodes hosts 2 independent sources, so that the number of sources is 50. The traffic of these TCP sources is aggregated in the Lyon switch (see Figure 2). Since this aggregated traffic has an average rate smaller than 1 Gbps, there is no congestion at the output port of the Lyon switch. This aggregated traffic shares the 10 Gbps bottleneck at the output of the Lyon router with a cross-traffic constituted of 20 permanent UDP flows from Lyon to Rennes, whose bandwidth (x Mbps) can be set to any value, thus allowing us to impose the loss rate.

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. All the TCP and UDP transfers are realized with *iperf* [10], a traffic generation tool which allows the users to tune the different TCP and UDP parameters. On the nodes running TCP flows, we collect *netstat* statistics to compute the loss rate.

During the experiments, the output traffic of the Lyon switch (corresponding to the aggregation of the 50 TCP sources) is captured. The capture device is based on the use of GtrcNET-1 [11], a FPGA based hardware that extracts the packet headers and forwards them to a storage server.

C. Data processing and flow reconstruction

To go from the packet level trace captured at the output of the Lyon Switch to the aggregated traffic we want to analyse, we use an original tool which features two major functions:

- *Packet aggregation*: grouping and counting packets in each contiguous time interval of size Δ yields the aggregated traffic time series $X^{(\Delta)}(t)$. In our work, we chose $\Delta = 10$ ms.
- *Flow reconstruction*: recomposing each flow from the intertwined packet stream is a delicate task. In our tool, a flow is classically defined as a set of packets sharing the same source and destination IP addresses, source and destination ports, and the same protocol. In addition, a `timeout` threshold is set so that two packets separated by a time larger than `timeout` cannot pertain to the same flow. This `timeout` can be used to detect silent times larger than `timeout` within a flow. It can take any value, including infinity.

Thanks to this tool, we can extract the size of each flow (and then the flow size distribution) and the aggregate traffic time series from the captured trace.

D. Experiments description

The number of independent sources in our experiments is 50. In all our experiments, the ON times are heavy-tailed distributed with tail index $\alpha = 1.5$ and the mean flow size is $1,000$ packets. The OFF times are exponentially distributed, with mean $\mu_{OFF} = 0.6$ s.

We used the same successions of flow sizes and OFF times corresponding to these distributions in three different experiments with different loss rates imposed by the UDP cross-traffic. Table I summarizes the loss rates achieved and the corresponding sending rate for the UDP traffic.

The first experiment is performed without any UDP cross-traffic to avoid packet loss. In this experiment, a TCP window limitation is used 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 congestion control mechanism to an average value depending on the loss rate. Table I shows the mean ON and OFF times really measured from the traces, where the flows were reconstructed with an infinite `timeout`.

IV. Results

A. LD of the aggregated traffic

To study the scaling laws of the aggregated traffic time series $X^{(\Delta)}(t)$, we use a wavelet based tool [12] which is

	UDP flow 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 I
EXPERIMENT PARAMETERS

	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 II
ESTIMATED LRD INDICES

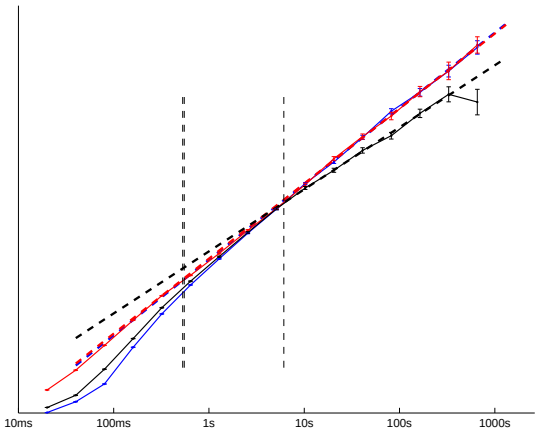

Fig. 3. Log-Diagrams for the three experiments with $\alpha = 1.5$: (bleu) no loss – (red) loss rate 0.7% – (black) loss rate 5%. The vertical lines materializes the mean ON times as shown in Table I.

known to perform very well. It is based on the log-diagram (LD) which represents the variance of the wavelet coefficients against the scale. For LRD signals, the LD is linear in the large scales, and the estimation of H then reduces to a linear regression.

Figure 3 shows the LDs of the three experiments with $\alpha = 1.5$. The mean ON time for each experiment has been materialized because it has been identified in [13], [6] as the minimal scale beyond which LRD should be observe as predicted by Taqqu's theorem.

As we want to investigate Taqqu's prediction, we now focus on the large time scales (beyond the mean ON time). Figure 3 exhibits for the three experiments a linear behavior of the LD in the large time scales. It means that there exists a scaling behavior in the large time scales, which is then likely to be related to the LRD prediction of Taqqu's Theorem. However, Figure 3 also shows a clear distinction between the first two experiments (no loss and loss rate 0.7%) which have very similar large scale scaling indices; and the third experiment (loss rate 5%) where the large scale scaling index is clearly smaller. Table II, which summarizes the estimated values of the large scale scaling indices, shows that the estimated value of H for the first two experiments are very close to the value of Taqqu's prediction ($H = 0.75$); whereas the estimated value of H for the third experiment is very close to 0.5 (no LRD). It seems then that a sufficiently high loss rate is able to annihilate the LRD of the aggregated traffic.

B. Interpretation with Taqqu's theorem

We now propose an interpretation of the previously stated observation that the LRD of the aggregated traffic induced by the heavy-tailed distributed ON times, can disappear under a sufficiently high loss rate, which is fully coherent with Taqqu's Theorem. The key point of this interpretation is the association of an ON period of the ON/OFF model with the duration of a flow, which can strongly depend on the `timeout` used for the flow reconstruction.

Figures 4, 5 and 6 display 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. It then 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, Fig. 4 and loss rate 0.7%, Fig. 5), the distributions reconstructed with the two values of `timeout` are almost the same. It means that almost no flow is experiencing an internal gap of more than 100 ms. This is the normal case for the no loss experiment since the packets of a flows are emitted by bursts separated by one RTT. For a small loss rate of 0.7%, it shows that the TCP window is rarely falling 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% (Figure 6), the distributions reconstructed with the two values of `timeout` are very different. The flow size distribution, which appears heavy-tailed with an infinite `timeout`, becomes exponential when the flows are reconstructed with a `timeout` of 100 ms. It means that the flows are experiencing internal gaps of length larger than 100 ms. These gaps correspond to the intervals between two packet retransmission during the exponential backoff periods, whose sizes take the values $2^k \times RTO_0$, $k = 1, 2, \ldots$ (with RTO_0 almost equal to $4RTT$, corresponding to about 200 ms with congestion). This is fully consistent with the observed OFF time distributions (Figure 6): with a `timeout` of 100 ms, it exhibits, in addition of the imposed OFF times (observed when reconstructing the flows with an infinite `timeout`), large peaks around the values 200 ms, 400 ms and 800 ms. These peaks correspond to the typical intervals between packet retransmission during exponential backoff periods: 1×200 ms, 2×200 ms and 4×200 ms respectively. The reconstruction of the flows with a `timeout` smaller than the mean OFF time then evidences the presence of internal gaps inside the flows reconstructed with an infinite `timeout`. We believe that, as these internal

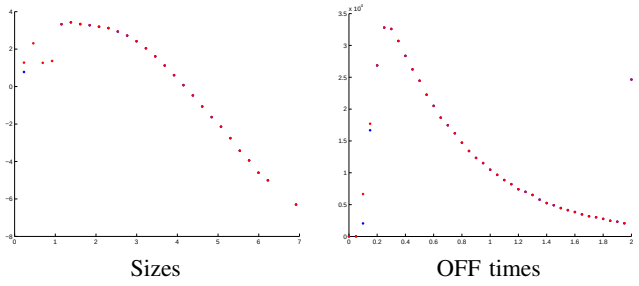

Fig. 4. Flow size distribution in log-log (left) and OFF time distribution (right) for the experiment with no loss.

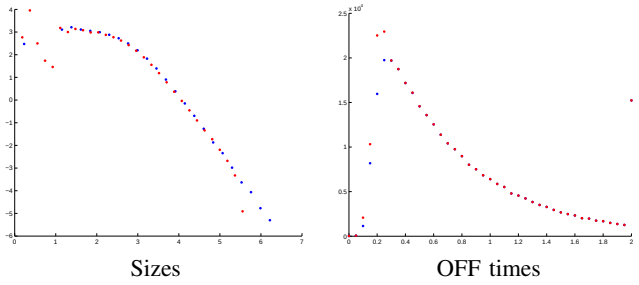

Fig. 5. Flow size distribution in log-log (left) and OFF time distribution (right) for the experiment with loss rate 0.7%

gaps 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, we have to use the flows reconstructed with a `timeout` of 100 ms to use the ON/OFF model and interpret the LRD index using Taqqu's Theorem. Indeed, in our last experiment with loss rate 5%, as the flow size distribution with a `timeout` of 100 ms is exponential, Taqqu's theorem predicts that there should be no LRD ($H = 0.5$), which we actually observe. We believe that this gives a limit to the interpretation of a source behavior as a ON/OFF source: packets should be "regularly" arriving during a ON period with no internal gap of the order of magnitude of the OFF times. Practically, to respect this limitation, one has to use a `timeout` for the flow reconstruction smaller than the mean OFF time.

V. Conclusion and future works

In this work, we showed that the LRD property of the aggregated traffic caused by the heavy-tailed flow size distribution, which is observe with a low loss rate, can disappear under a sufficiently high loss rate. We showed that this is not in contradiction with Taqqu's Theorem which predicts the observation of LRD in the aggregated traffic in an ON/OFF model with many sources when the ON time distribution is heavy-tailed. It is due to the exponential backoff phases (coming from the TCP congestion control mechanism), which are of the same order of magnitude as the mean OFF time, and thus makes it impossible to associate a flow with a single ON time. Instead, the flows have to be split into several smaller flows, thus making the new flow size distribution exponential. The absence of the LRD property in this heavy loss case is then fully consistent with Taqqu's prediction.

In a future work, we plan to investigate further the LRD of aggregated traffic with congestion under different points: with different mean OFF times and RTTs, and with heavy-tailed distributed OFF times. We also plan to investigate from both a theoretical and an experimental viewpoint the new flow size distribution, and the parameters influencing the limit loss rate which separate the case where the flow size distribution remains heavy-tailed from the case where it becomes exponential.

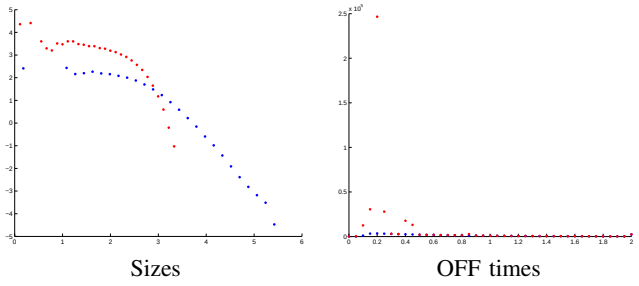

Fig. 6. Flow size distribution in log-log (left) and OFF time distribution (right) for the experiment with loss rate 5%

References

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

[2] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *ACM/IEEE Trans. on Net.*, vol. 2, no. 1, pp. 1–15, Feb. 1994.

[3] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and possible causes," *IEEE/ACM Trans. on Net.*, vol. 5, no. 6, pp. 835–846, Dec. 1997. [Online]. Available: http://www.cs.bu.edu/faculty/crovella/paper-archive/self-sim/journal-version.pdf

[4] M. S. Taqqu, W. Willinger, and R. Sherman, "Proof of a fundamental result in self-similar traffic modeling," *SIGCOMM CCR*, vol. 27, no. 2, pp. 5–23, 1997.

[5] D. R. Figueiredo, B. Liu, A. Feldmann, V. Misra, D. Towsley, and W. Willinger, "On TCP and self-similar traffic," *Performance Evaluation*, vol. 61, no. 2-3, pp. 129–141, 2005.

[6] P. Loiseau, P. Gonçalves, G. Dewaele, P. Borgnat, P. Abry, , and P. Vicat-Blanc Primet, "Investigating self-similarity and heavy-tailed distributions on a large scale experimental facility," INRIA, Tech. Rep. 6472, March 2008.

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

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

[9] R. Bolze, F. Cappello, E. Caron, M. Daydé, F. Desprez, E. Jeannot, Y. Jégou, S. Lanteri, J. Leduc, N. Melab, G. Mornet, R. Namyst, P. Primet, B. Quetier, O. Richard, E.-G. Talbi, and I. Touche, "Grid'5000: a large scale and highly reconfigurable experimental grid testbed." *Int. J. of High Performance Computing Applications*, vol. 20, no. 4, pp. 481–494, nov 2006.

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

[11] Y. Kodama, T. Kudoh, R. Takano, H. Sato, O. Tatebe, and S. Sekiguchi, "Gnet-1: Gigabit ethernet network testbed," in *Proc. of 2004 IEEE International Conference on Cluster Computing (Cluster2004)*, 2004, pp. 185 – 192.

[12] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch, "Wavelets for the analysis, estimation and synthesis of scaling data," in *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds. John Wiley & Sons, Inc., 2000.

[13] N. Hohn, D. Veitch, and P. Abry, "Cluster processes, a natural language for network traffic," *IEEE Trans. on Sig. Proc. – Special Issue on Sig. Proc. in Net.*, vol. 8, no. 51, pp. 2229–2244, Oct. 2003.