# Metroflux: A fully operational high speed metrology platform

Patrick Loiseau, Paulo Gonçalves, Yuetsu Kodama, Pascale Primet Vicat-Blanc

## I. Motivations

Modern experimental research works on networks are facing some difficulties due to the emergence of very high speed links (10 Gbps). Performing realistic experiments, in order to relevantly characterize the network traffic then requires to solve the challenging issue of monitoring these very high speed links at a fine grain (packet level).

On the other hand, while real production link traffic analysis remains a central problem, we also need to perform fully controlled and reproducible experiments which allows the user to variate a lot of parameters (aggregation level, congestion, source flow size distribution, etc.); and then study separately their impact on the traffic.

In this paper, we present Metroflux, a fully operational metrology platform, based on the GtrcNet-1 and GtrcNET-10 hardware, which enables to capture the traffic at packet level on a very high speed link and to analyze it at flow level. We then present its utilization in two different situations: a controlled experiment on Grid500 and the monitoring of a production link.

## II. General design of the Metroflux system

### A. Global architecture of the system

Metroflux is a programmable system for flow analysis which currently operates on a 1 Gbps link without loss, and which also support 10 Gbps links. It is composed of hardware and software. The system is able to monitor a 1 Gbps or a 10 Gbps link, by capturing without loss the first bytes of every packets, and then saving them to a pcap file (this file format is used by tcpdump and saves packets with their capturing timestamp). The duration of the capture depends on the packet size and the throughput of the link. The Metroflux system integrate the GtrcNet-1 box (http://projects.gtrc.aist.go.jp/gnet/gnet1e.html) or the GtrcNet-10 box (for 10 Gbps links) and a storage server with large amount of disk space. The server runs a packet header extracting program, which is based on the MAPI library (http://mapi.uninett.no/, Monitoring API developed by the LOBSTER project, http://www.ist-lobster.org/). MAPI provides functions to manipulate packet headers in pcap format.

In addition, a original and extensible library for traffic and flow statistical analysis has been developed along with a tool to design a traffic analyze experiment and to automatically deploy it.

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)

Yuetsu Kodama is with National Institute of Advanced Industrial Science and Technology (AIST), Japan. (e-mail: y-kodama@aist.go.jp)

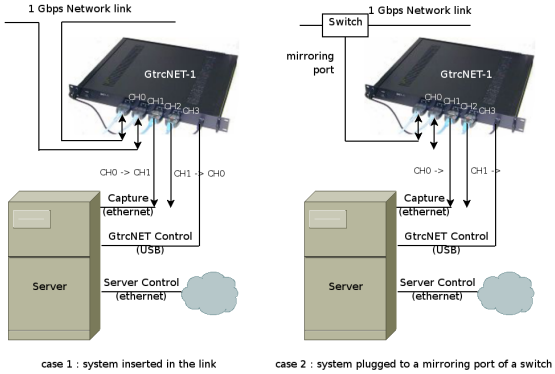


Fig. 1. Two different ways of installing Metroflux on a link

The Metroflux system can be installed transparently within an experimental or a production network in two different ways (see Figure 1):

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

### B. GtrcNET-10 characteristics for 10 Gbps links measurement

GtrcNET-10 [1] is a hardware layer 2 network equipment with 10 GbE ports, and it is a successor of GtrcNET-1 [2] with GbE ports. GtrcNET-10 provides many parametrable functions, such as traffic monitoring in microsecond resolution, traffic shaping, and WAN emulation at 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 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. Fig. 2 shows the architecture of GtrcNET-10. The FPGA is a Xilinx XC2VP100, which includes three 10 Gbps Ethernet MAC and XAUI interfaces. The FPGA can directly receive the 3.125 GHz signals from XENPAK module by Rocket I/O hardware macro. The main circuit in FPGA runs on 156.25 MHz clock with 64 bit data width, and the memory is accessed by 162 MHz to support read and write with wire rate speed of 10 GbE. GtrcNET-10 is connected to a control PC via USB. It also has a serial port to connect GPS module, and it can synchronize the time

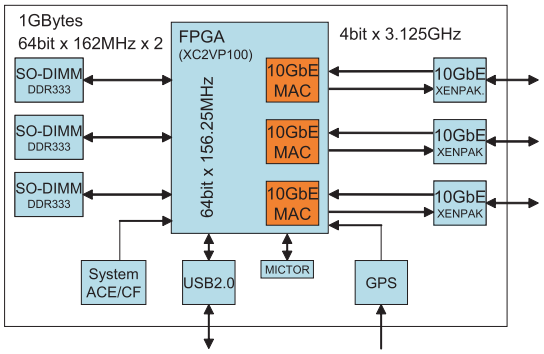


Fig. 2. The architecture of GtrcNET-10

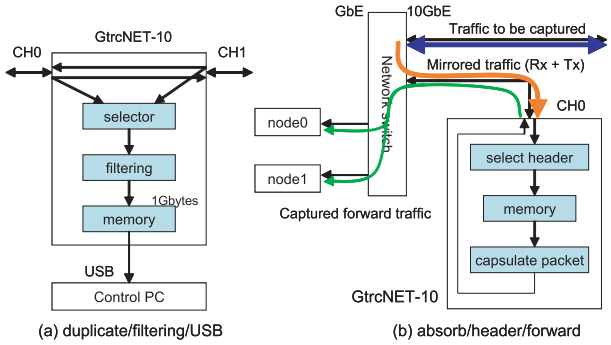


Fig. 3. Example of packet capture on GtrcNET-10

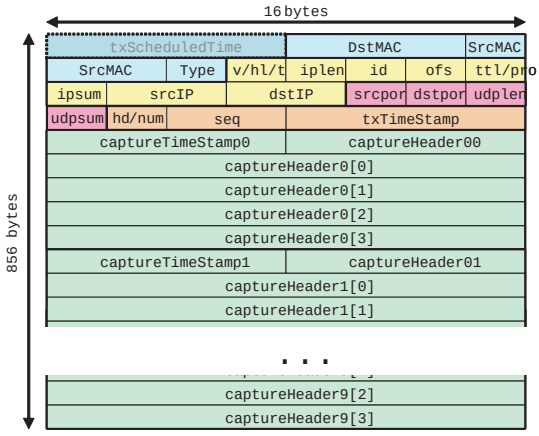


Fig. 4. Packet format of capture forward

with other GtrcNET by receiving accurate time from GPS. The MICTOR connector is for debug in the circuit of FPGA. System ACE initializes the FPGA by the configuration data on Compact Flush memory.

By programming FPGA, one can add new functions and improve existing functions according to the requirements of the users.

Let us now detail packet capture functions of GtrcNET-10 with two examples.

Fig. 3 (a) shows an example of usage for packet capture on GtrcNET-10. Between CH0 and CH1, traffic to be captured is transferred. GtrcNET-10 selects an input port, duplicates packets from the port, and captures them in a memory. This function cannot capture bi-directional traffic. Traffic between CH0 and CH1 is not affected by the capturing. Packets are captured only if they satisfy the filtering conditions. A condition indicates a 16 bit field of header, selects any bits by 16 bit mask, and compares it with the specified 16 bit value. One or two conditions are specified, and the logical combination of two conditions (logical-and or logical-or) is also specified. For example, you can capture only packets whose source port or destination port of TCP/IP header coincides with a specified value. Also you can capture only packets that have VLAN tag and the VLAN ID is a specified value. The captured data is read from a control PC via USB. Since the USB access speed of 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 10 GbE are captured, only 800 ms traffic can be captured.

Fig. 3 (b) shows another example of usage for packet capture on GtrcNET-10. Traffic to be captured is mirrored by a network switch, and the mirrored traffic transferred to GtrcNET-10. GtrcNET-10 captured only the selected header fields of packets in a memory. The selection of header is 16 bytes field, and any field can be selected independently up to 128 bytes. The received timestamp, which is 64 bit and the format is based on RFC-1305 [3], and first 8 bytes of header are always captured. Multiple captured headers are encapsulated into a UDP packet, and transferred from an output port. Fig. 4 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 follows 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 Fig. 4, the capture forward traffic rate is about 556 Mbps. The forward traffic should be read by a receive node. If the traffic is too large to read by single node, the destination of the forward traffic can be distributed to multiple nodes in round-robin, up to 16 nodes.

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 for packet capturing of one port.

GtrcNET-10 can capture packets in any combination, such as duplicating packets or not, filtering packets or not, selecting header or not, accessing by USB or forwarding as packets, but capture forwarding is only for selected headers.

In Metroflux system, GtrcNET-10 is used with a server

which stores the packet headers' stream and perform off-line fine grain traffic analysis.

### C. Overview of the advantages of Metroflux

There are not yet a lot of solutions on the market to perform network traffic capture at speeds as high as 10 Gbps. Among the available solutions, our Metroflux platform has some distinctive features:

- It only captures packet headers, whereas other solutions often capture full traffic. Currently, packet headers (Ethernet + IP + TCP/UDP) are enough for us to study the traffic.
- Being base on a programmable FPGA network appliance separated from the capture server (instead of an integrated capture card) has some strong benefits:
  - It allows a lot of processing to be performed in hardware, including filtering and/or sampling of traffic to capture.
  - Moreover, the GtrcNET-10 is able to perform other functions (latency emulation, accurate bandwidth measurements, traffic shaping) and multiple functions can be performed simultaneously.
  - The functions of the GtrcNET-10 can be expanded as needed (provided one has the FPGA development expertise).
  - We have full control on the whole packet capture chain.

  The drawback of this solution is the added complexity to build, setup and operate the Metroflux platform, compared to ready-made solutions that can be found on the market.
- A single GtrcNET10 can distribute its load among several capture servers

## III. Example I: a fully controlled experiment on Grid5000

As a first example of the utilization of Metroflux, we present a fully controlled experiment performed on Grid5000. The aim of this experiment is to validate on a large scale experimental network the link between the tail index of the (heavy-tailed) flow size distribution and the long-range dependence (LRD) parameter of the aggregated traffic [4], [5].

Note that this experiment was initially performed with a 1 Gbps bottleneck and a GtrcNET-1 was used. However, the method remains the same when using a GtrcNET-10 on a 10 Gbps link.

### A. Grid5000

Grid5000 is a 5000 CPUs nation wide grid infrastructure dedicated to network and grid computing research [6]. Up to 17 French laboratories are involved and 9 sites geographically distributed are hosting one or more cluster of about 500 cores each (Fig. 5). The sites are interconnected by a dedicated optical network provided by RENATER, the French National Research and Education Network. It is composed of private 10 Gbps Ethernet links connected to a DWDM core with

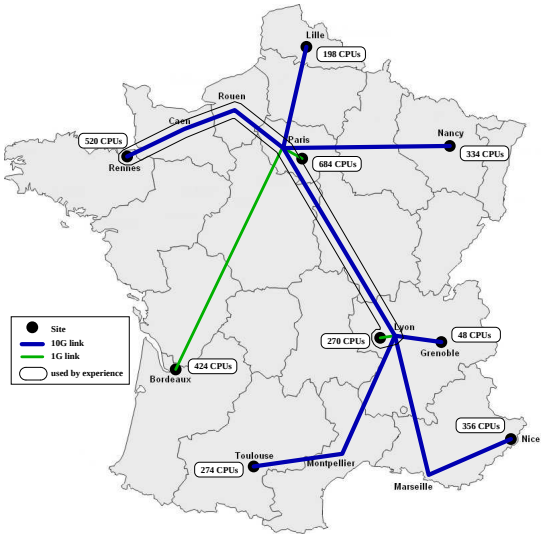


Fig. 5. Grid5000

dedicated 10 Gbps lambdas, with a bottleneck at 1 Gbps in Bordeaux and Lyon. Two international interconnections are also available: one at 10 Gbps with DAS3 (the Netherlands) and one at 1 Gbps with Nareji (Japan).

Grid5000 is a research tool, featured with deep control, reconfiguration and monitoring capabilities to complement network simulators and emulators. It allows the users to reserve the same set of dedicated nodes across successive experiments, and to have full controll of these nodes to run their own experimental condition injector and measurement software. This can be achieve thanks to two tools: OAR and Kadeploy. OAR is a reservation tool which offers advanced features (CPU/Core/Switch reservation). Kadeploy is an environment deployment system which allows the users to have their own customized environment automatically deployed on a large number of nodes. For example, kernel modules for rate limitation, congestion control variants or QoS measurement can be added to the native operative system.

Finally, as a private testbed dedicated to research, Grid5000 makes easy the installation of experimental hardware like Metroflux at representative traffic aggregation points. In the scenario detailled bellow, the Metroflux system is measuring the access link of the Grid5000 site at Lyon.

### B. Scenario description

The topology used for this experiment is described on Figure 6: 100 clients nodes in Lyon (sources) are emitting to 100 server nodes in Rennes (destinations). The average $RTT$ is 12 ms. Each client behaves like a ON/OFF source: ON periods correspond to a flow emission and OFF periods to an idle time separating two consecutive ON periods. The source rate during an ON period is limited to 5 Mbps to avoid congestion at the 1 Gbps bottleneck. The rate limitation mechanism can be chosen: pspacer, token bucket or TCP window limitation.

For the purpose of our experiment, the ON periods are independent and identically distributed random variables, fol-

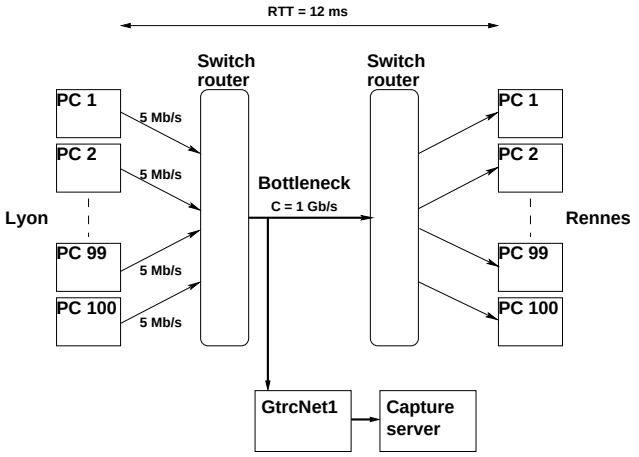


Fig. 6. Experimental topology

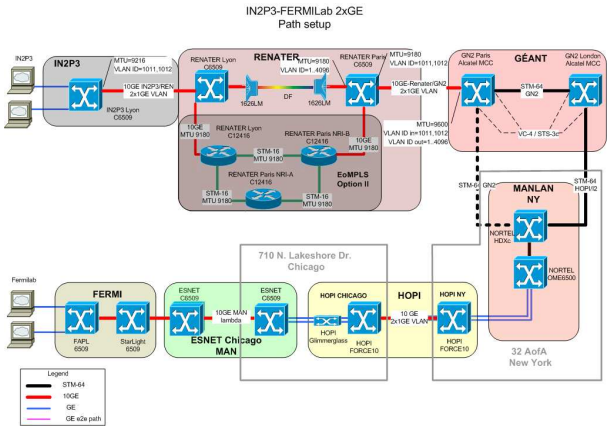


Fig. 7. in2p3-fermilab link

lowing a heavy-tailed distribution of tail index $\alpha$. The OFF periods are exponentially distributed with the same mean as the ON periods.

During the 8-hours experiment, the out-going traffic from Lyon is mirrored and sent to the capture device (see Fig. 6). It results in a trace stored on the server, containing all the packet headers with the associated packet timestamps, which can be used for off-line statistical analysis.

### C. Analysis and results from the captured trace

All the informations needed to compute the tail index and LRD index can now be retrieved from the trace.

Firstly, grouping and counting the packets in each contiguous time interval yields the aggregate traffic time series, from which the LRD index can be measured using appropriate wavelet-based tools [7].

Secondly, flows can be reconstructed from the trace. To do so, the packets sharing the same IP source and destination addresses, the same source and destination ports and the same protocol are grouped into a flow, as long as two such consecutive packets are not separated by more than a threshold named `timeout`, whose value has to be carefully chosen. From this flow reconstruction, we then can easily extract the flow sizes, and then estimate the tail index to verify that it is the same as the one imposed for ON periods. The tail index estimator used is a recent wavelet-based estimator whose good qualities are shown in [8].

Thus, from a trace capture at an aggregated link, we are able to measure the tail index and the LRD index, and then to check the link between these two indices. The use of the fully controllable tool Grid5000 allows us to perform the same experiment many times, while varying a lot of parameters: the protocol, the tail index, the rate limitation, etc. The impact of all these parameters can then be thoroughly studied. More details about this kind of experiments, and the results can be found in [9].

### D. Sampling

Even if GtrcNET-10 enables full packet capture at 10 Gbps, the capture of all the timestamped packet headers at this speed can prove very demanding in terms of storage resources. To limit this resource consumption, a sampling function has been implemented in GtrcNET-10: when activated, it captures only one packet every $N$ packets, for an arbitrary value of $N$.

Although, because of its aggregated characteristic, the LRD parameter can still easily be estimated from a sub-sampled trace, the estimation of the tail index becomes much more complicated in this situation. To solve this problem, we have derived in [10] the exact maximum likelihood solution for the estimation of the tail index from sub-sampled data.

## IV. Example II: Analysis of Real Production Grid Traffic

### A. Description of the analyzed link

Figure 7 shows the paths of the production traffic between in2p3 (Lyon, France) and fermilab (Chicago, USA). We "plugged" Metroflux at the output of the in2p3 to capture the traffic of the 1 Gbps Ethernet VLAN encapsulated in the 10 Gbps link, using a GtrcNET-1.

### B. Results

We now present the results obtained from a 50 days continuous capture on the link described in the previous subsection. Although we were able to monitor outgoing traffic as well as ingoing traffic, we only present the results associated with ingoing traffic (both have the same characteristics).

As explained in Example I, we are able to reconstruct from the capture the list of all the flows with their characteristics (length (in packets), volume (in Bytes), duration, etc.). Figure 8 shows the distribution of the flow volume and duration, where we separated three different types of flows with respect to the mean size $\mu$ of the flow's packets: $\mu = 64$ (small packets, mainly ACKs' flows), $\mu = 1448$ (large packets, typical of data's flows), $64 < \mu < 1448$ (intermediate). Interestingly, we can see that the tail of the flow volume distribution is constituted by flows of large packets (blue curve), but the tail of the flow duration distribution is constituted by flows of small and intermediate packets (red and green curves). It means that the largest flows in term of volume are not the longest flows in term of duration.

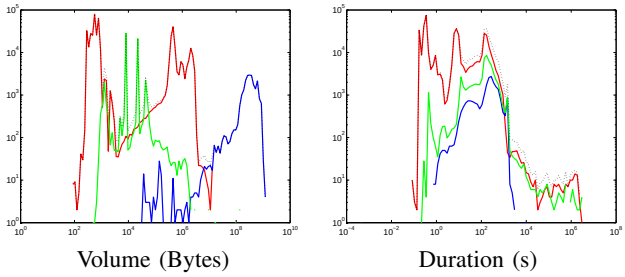


Fig. 8. Distributions (loglog plot) of the flow volume and duration for different mean packet size. ($\mu$, in Bytes): (red) $\mu = 64$ – (blue) $\mu = 1448$ – (green) $64 < \mu < 1448$

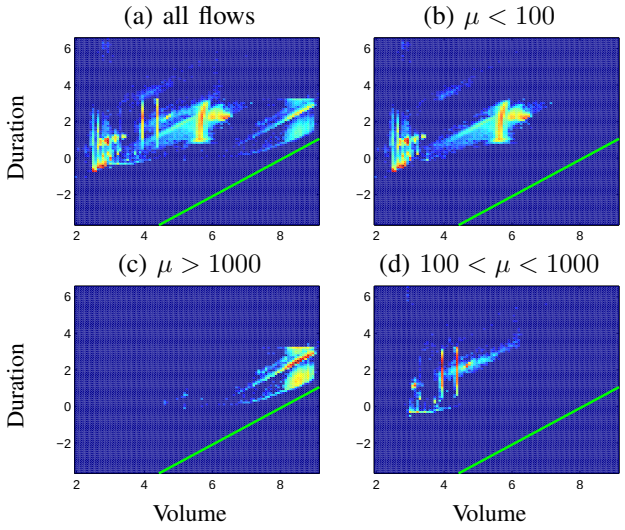


Fig. 9. Flow duration against flow volume for different mean packet sizes. The line is the minimal flow duration for a given volume (at 1 Gbps).

To complement this study with a joint analysis of the duration and the volume, Fig. 9 represents the flow duration against the flow volume for the three flow types, and for all the flows. We see again in this Figure that the longest (in term of duration) flows are not the largest ones (in term of volume).

## V. Conclusion

In this paper we have presented Metroflux, a fully operational metrology platform based on the GtrcNET hardware. It enables a full packet header capture at very high speed links (10 Gbps) and a fine grain flow analysis. In two examples, we have shown how Metroflux can be used to monitor high speed links at packet level and get usefull insights on the characteristics of the traffic going through these links. It makes from Metroflux an original monitoring tool of primary importance for very high speed network research.

## VI. Acknowledgement

This work has been funded by the French ministry of Education and Research, INRIA, and CNRS, via ACI GRID's Grid'5000 project , the ANR IGTMD grant, IST EC-GIN project, INRIA GridNet-FJ grant, NEGST CNRS-JSP project.

## References

[1] http://www.gtrc.aist.go.jp/gnet.

[2] 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)*, pp. 185 – 192.

[3] D. L. Mills, "Network time protocol (version 3) specification, implementation and analysis," RFC 1035, March 1992.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] P. Gonçalves and R. Riedi, "Diverging moments and parameter estimation," *J. of American Stat. Asso.*, vol. 100, no. 472, pp. 1382–1393, December 2005.

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

[10] P. Loiseau, P. Gonçalves, S. Girard, F. Forbes, and P. Primet Vicat-Blanc, "Maximum likelihood estimation of the tail index of flow size distribution from sampled data," August 2008, preprint.