

# Results of Large-Scale Queueing Delay Tomography Performed in the ETOMIC Infrastructure

István Csabai, Péter Hága, Péter Mátray, Gábor Simon, József Stéger and Gábor Vattay  
Department of Physics of Complex Systems, Eötvös University, Budapest, Hungary  
Email: {csabai,haga,matray,gaba,steger,vattay}@complex.elte.hu

**Abstract**— This paper presents operational experience of large-scale unicast network tomography, that samples a part of the European Internet. In the paper we describe in detail the ETOMIC measurement platform that was used to conduct the experiments, and its potential in future scaled-up measurements. The main results of the paper are maps showing various spatial and temporal structure in the characteristics of queueing delay corresponding to the resolved part of the European Internet. These maps reveal that the average queueing delay on different network segments spans more than two orders of magnitude. At the most loaded time of day we find that the distribution of average queueing delays among the different segments follows closely a log-normal distribution.

## I. INTRODUCTION

The Internet is a huge network of computers and routers with highly heterogeneous properties, under decentralized administration. The measuring of the static and dynamical state of this complex network is a very important and inevitable task for predicting the quality of various services and applications over the Internet.

From the point of view of the current, most widely used data transfer protocols like the variants of TCP, the most relevant state variables of the network are the loss-rates and the delays encountered by data packets on a path, since these quantities control the transmission rate of the transfer protocol, that most applications rely upon. These characteristics also constitute the key metrics in the service level agreements of today's ISPs. An attractive way to measure packet loss-rate and delay over the Internet, is by means of active probing. Active probing involves injecting of probe packets into the network and analyzing the properties of the received probe-stream. This technique is flexible and has a wide range of applicability, however care must be taken not to overload the network by the extensive use of probe packets. In the past years several measurement platforms have been developed for conducting active measurements over the Internet (e.g. Surveyor, Felix, AMP [1]), however these platforms can provide only end-to-end information between the participating nodes, where the measured characteristic can not be resolved on the parts of the end-to-end path. Most of the existing methods giving resolved information on the parts of the end-to-end path, rely on extra cooperation of the routers in the path to process their packets. As the Internet continues to evolve towards more decentralized and heterogeneous administration, in the future the cooperation of the network elements can be foreseen to be limited to the basic process of just storing and forwarding incoming

probe packets. This trend motivates the development of novel measuring methods that do not rely on any responses of the routers.

The solution to the above problem is provided by network tomography, which is a special class of active-probing measuring techniques, that is able to resolve the end-to-end delay statistics [2], [3] and packet loss rates [4], [5] to internal segments of the paths. In general a tomography measurement made from a single source to a set of receivers admits the determination of the delay statistics and loss-rates on each segment of an underlying logical tree that is spanned by the source and receiver nodes, and the branching nodes (nodes where the path of probes originating from the same source, but destined for different receivers diverge). By increasing the number of sources and receivers involved in a measurement, in the limiting case the tomography approach resolves completely the network on true links, while end-to-end measurements do not share this property.

Initially network tomography techniques were developed for the use with multicast probes [6], [7], [5], [9], which requires the extra cooperation of the routers to support multicast functionality. These approaches were later also extended to the case of using unicast probes [2], [4], [5], and performing the measurements from multiple sources [10], [11], which makes unicast network tomography the most general tool to measure spatially resolved characteristics of an uncooperative network, like the Internet.

The main idea of unicast network tomography is to use back-to-back packet pairs, where each packet of a pair is destined to different receivers. As the packets of a pair traverse their paths, they experience the same network conditions on the common segment from the source to the branching node, which brings correlation into the time-series of the end-to-end characteristics. The correlation property of such unicast probe streams, which resembles multicast traffic, is the key to resolve the internal characteristics from the end-to-end measurements.

The delay experienced by a packet over an Internet path sums up from two non-negative components, a constant part (mainly due to propagation delay) and a time-varying component due to queueing in the buffers of routers. In this paper we are concerned with large-scale inference of queueing delay distributions in the Internet by performing extensive unicast network tomography measurements. The large-scale study of queueing delay distributions is motivated by the fact that this observable carries vast amount of information about

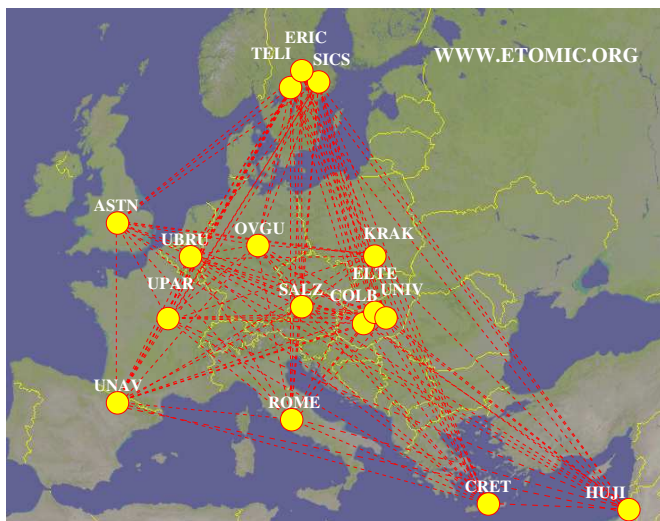


Fig. 1. The geographical locations of the deployed ETOMIC measurement nodes. The abbreviations of the nodes together with additional information are given in Table I.

traffic properties and the state of congestion on the measured path. By resolving the queueing delay distributions from end-to-end measurements we can draw a map of congestion of the network segments, analyze spatial structure, and identify highly congested or faulty segments. By performing such measurements repeatedly at different times of day admits also the study of temporal evolution of the network state. From another perspective being able to measure the distribution of queueing delays on large scale, in a heterogeneous real-world scenario, helps in the development of realistic Internet delay models, which is itself an active area of research.

The rest of the paper is organized as follows. In the next section we describe ETOMIC, the measurement platform, where the experiments were conducted. Afterwards we proceed with the presentation of the results of our large-scale tomography measurements, while the last section discusses the obtained results from several perspectives.

## II. THE ETOMIC MEASUREMENT PLATFORM

To perform unicast queueing delay tomography in a real network environment poses several challenges. First in order to be able to measure true end-to-end delay, source and receiver nodes need to be synchronized to a common clock-reference, and must stay in the synchronized state during the measurements. Second, the measuring infrastructure has to be very precise in order to be able to resolve the microsecond-scale queueing delay components associated to high-bandwidth (multi-Gigabit) links. The precision of commercial workstations with NTP synchronization are insufficient for this task, thus to achieve sub-microsecond precision, a hardware solution is inevitable.

In the subproject of the European Union sponsored EVERGROW Integrated Project[12], we are developing a state of the art high-precision, synchronized measurement platform, the Evergrow Traffic Observatory Measurement InfrastruCture

TABLE I  
LIST OF CURRENT ETOMIC MEASUREMENT NODES.

Abbreviation	IP address	Location
SICS	193.10.64.81	Stockholm, Sweden
TELI	217.209.228.122	Stockholm, Sweden
ERIC	192.71.20.150	Stockholm, Sweden
UNAV	130.206.163.165	Pamplona, Spain
ASTN	134.151.158.18	Birmingham, England
HUJI	132.65.240.105	Jerusalem, Israel
OVGU	141.44.40.50	Magdeburg, Germany
ROME	141.108.20.7	Rome, Italy
UNIV	193.6.205.10	Budapest, Hungary
COLB	193.6.20.240	Budapest, Hungary
ELTE	157.181.172.74	Budapest, Hungary
UPAR	193.55.15.203	Paris, France
SALZ	212.183.10.184	Salzburg, Austria
UBRU	193.190.247.240	Brussels, Belgium
CRET	147.27.14.7	Chania, Greece
KRAK	149.156.203.242	Krakow, Poland

(ETOMIC)[13]. This platform among others provides the ability to perform large-scale queueing-delay and loss tomography based on unicast probing techniques, and will be generally available and open to the public. Currently ETOMIC consists of 16 measuring nodes deployed at different locations in various European countries (See Fig 1 and Table I for details), while continuous efforts are made to incorporate new nodes into the system every year. The measurement nodes and the network experiments are managed through a central management system that is accessible to the researchers through a web-based graphical user interface [14], [15]. An ETOMIC measurement node is basically a standard PC (see Fig. 2a), but which in addition to its standard network interface card also includes an Endace DAG 3.6GE card, that is specifically designed for precise active and passive measurements [16]. These DAG cards provide very accurate time-stamping of the probe packets, with a time-resolution of 60 ns, and also advanced capabilities for transmission. A burst composed of several packets can be transmitted with precise user-defined inter-packet timings. In addition to the above, all the ETOMIC measuring nodes are connected to a GPS (Garmin GPS 35 HVS), that provides a PPS (pulse per second) reference signal directly to the DAG card. Thus global synchronization of the nodes can be achieved.

The accuracy of one-way delay measurements between two ETOMIC nodes is found to be  $\approx 0.5 \mu s$ , which is mainly limited by the performance of the GPS receivers. This result was obtained in a lab experiment in advance of the deployment of the nodes. In this experiment we connected two ETOMIC nodes by an empty link, and transferred a long stream of probe packets between them. Figure 2b shows the histogram of the measured delay, where the bin size reflects the time resolution of the DAG-cards. By performing extensive traceroute measurements between the available ETOMIC nodes we are able to determine the connection topology. For example a measurement that involved 10 ETOMIC nodes (both as sources and as destinations), the resulting connection topology contained 51 branching nodes and 130 network segments, among which many are true links in the GÉANT multi-

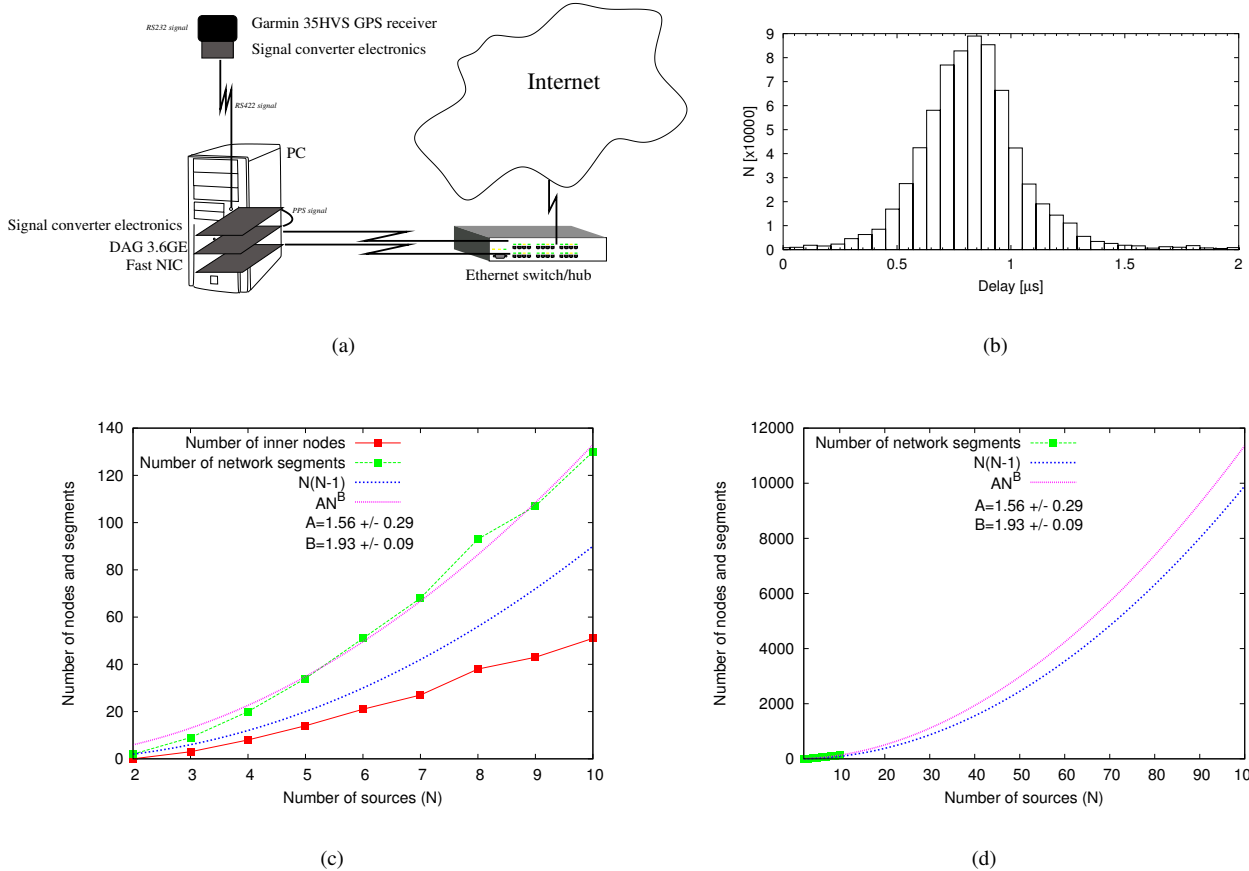


Fig. 2. (a) The schematics of an ETOMIC measurement node. (b) Histogram of measured one-way delay between two such nodes over an empty link indicating a precision of  $\approx 0.5\mu$ s. (c) The scaling of the number of network segments (edges) and branching (inner) nodes with the number ETOMIC source nodes. Finally (d) shows the extrapolated scaling of the number of network segments.

Gigabit European academic network [17] with link speeds ranging from 2 to 15 Gb/s. By reconstruction of the connection topology after dropping more and more sources from this data we can study how the number of measurable network segments and branching nodes scales with the number of sources.

Figure 2(c) shows the results, where we also indicate the scaling of end-to-end paths that asymptotically converges to  $\approx N^2$ . As can be seen the number of resolved network segments is significantly higher than the number of end-to-end paths. Assuming a power-law relation fitted to the data admits the extrapolation of the scaling for larger system sizes (see Fig. 2(d)). According to the fit, by doubling the number of current ETOMIC nodes would admit the resolution of  $\approx 1000$  network segments. Note however, that for large source numbers the data expected to deviate from the fit and saturate at the number of true links in the studied region of the Internet, while the number of end-to-end paths would still scale as  $\approx N^2$ .

### III. LARGE-SCALE QUEUEING DELAY TOMOGRAPHY

In this section we describe the large-scale tomography measurements performed at three different times of day (03:30, 14:05, 16:12), that correspond to different levels of congestion in the network. These measurements involved 9 ETOMIC

nodes, and a connectivity graph consisting of 38 branching nodes and 93 network segments. The branching nodes and network segments are embedded in the real Internet, where the network elements are administrated by different organizations. Among these 9 ETOMIC nodes 8 were simultaneously sources of outgoing back-to-back probe pairs, and receivers of the incoming probe packets, while one of the nodes was only used as a receiver. Each of the source nodes sent probe pairs consisting of 40Byte UDP packets to all the possible pairs of receivers in a round-robin fashion with an inter-pair time of 1 ms, and repeated this process many times. This procedure finally resulted in data sets, each containing two correlated time-series of end-to-end delays with an approximate length of 10000 elements. The end-to-end queueing delays can easily be calculated with the subtraction of the fixed delays (propagation, etc.). The sum of these fixed delays can be identified as the well defined minima of the one-way delay data. Before and after the packet-pair measurements we also performed connection topology detection with traceroute to determine possible changes during the measuring process.

These data sets comprised the input to the tomography method, that yielded the queueing delay distributions resolved

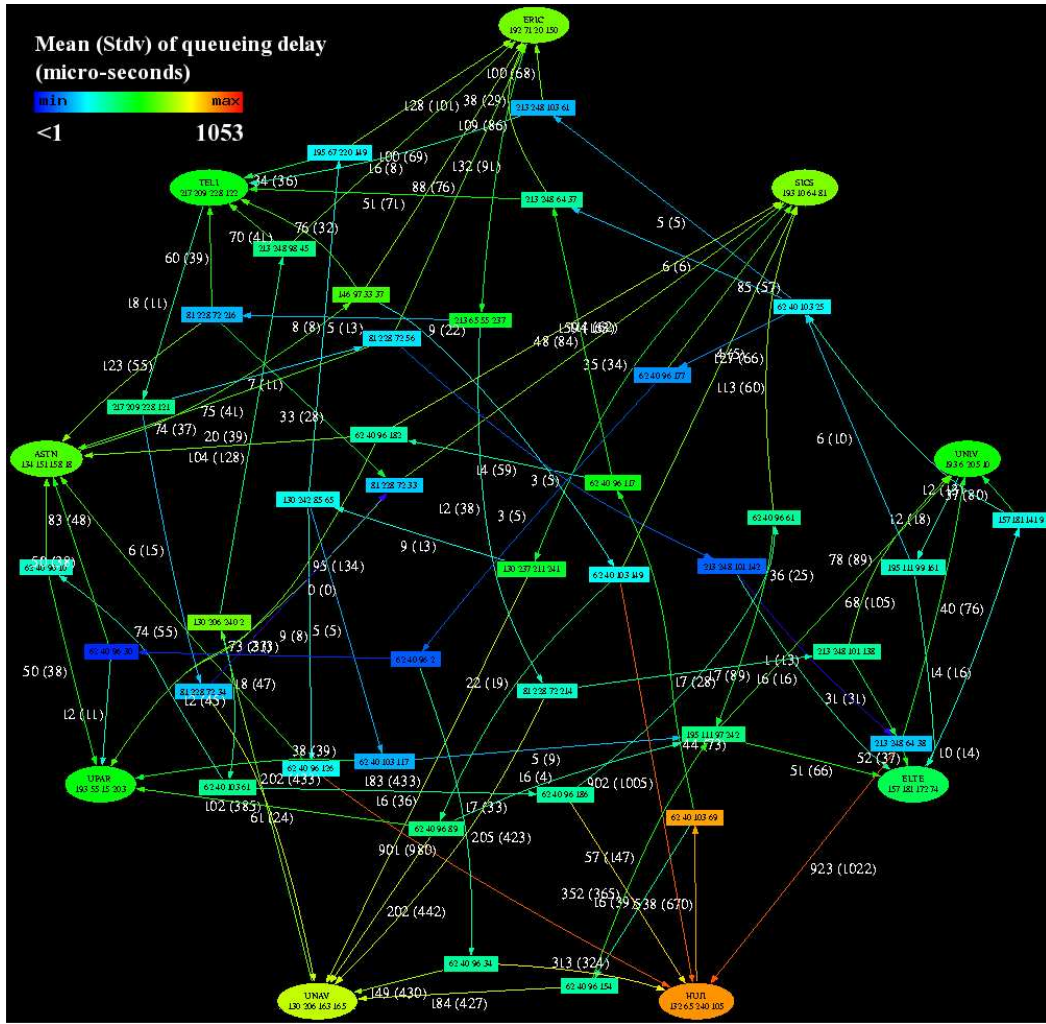


Fig. 3. The connectivity graph colored by the mean and labelled by the mean and the standard deviation (in brackets) of the queueing delay, given in units of  $\mu\text{s}$ . The ellipse shaped nodes on the edge are ETOMIC measurement nodes with abbreviations given in Table I, while box-shaped nodes in the interior of the graph are branching nodes. The arrows indicate the direction of probe packet flow on a given network segment. The boxes with an IP address of 62.40.X.X are nodes within GÉANT.

for each segment contained in the connectivity graph as an output. Since a given segment can be a part of different end-to-end paths, this fact enabled to test the consistency of the results, as well as the averaging of the distributions obtained from different data sets, but attributed for the same segment. The method we used to infer the queueing delay distributions is based on the quantization of the measured end-to-end delays into  $10\mu\text{s}$  bins, maximum-likelihood estimation via the expectation maximization algorithm, and algebraic deconvolution via the non-negative least squares algorithm. This method is described in detail in references [18], [19] that also provide the derivation and performance evaluations in ns-2 simulations and controlled lab experiments.

For better visualization of the results we extracted the mean and the standard deviation of the queueing delay distributions. The results for the mean in the case of the most loaded time of day are shown in Fig. 3, while the rest of the data along

with all of the queueing delay distributions can be accessed from the ETOMIC web-page by following the *Visualization* link.

#### IV. ANALYSIS OF THE RESULTS

The results of Fig. 3 reveal some interesting structure. First of all the mean of the queueing delay on the different segments spans three orders of magnitude, ranging from the error limit of the delay measurements ( $\approx 0.5\mu\text{s}$ ), to an average queueing delay of  $\approx 1$  ms, that characterizes a segment which connects GÉANT to the Hebrew University in Jerusalem. The results also reveal an interesting geographical feature, namely that the segments originating or ending in ETOMIC nodes that are located on the south (HUJI, UNAV), are characterized by the highest average and standard deviation of the queueing delays. As a general feature it can be observed that for all end-nodes incoming segments are characterized by higher

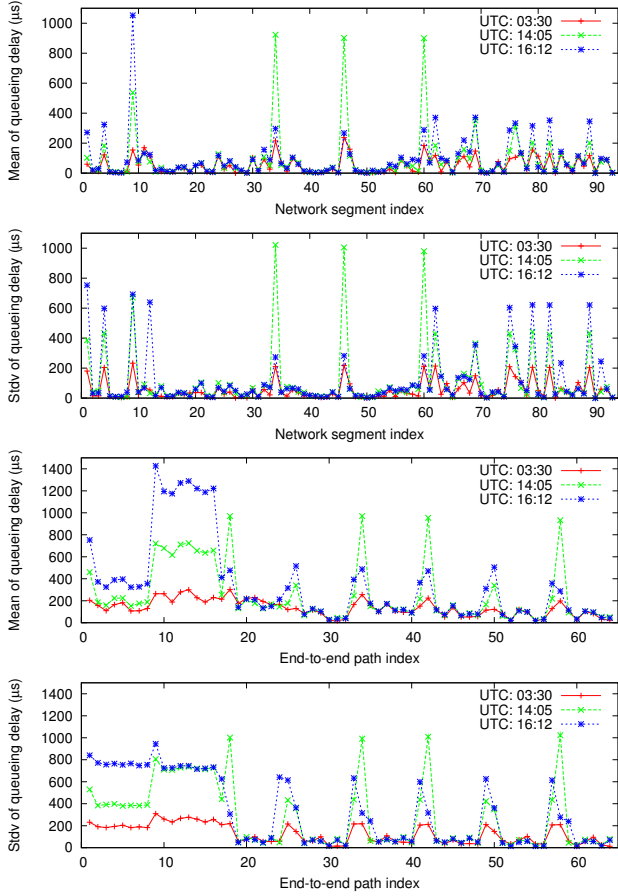


Fig. 4. Temporal variation of the mean and standard deviation of queuing delays for all the different end-to-end paths, and for all the network segments involved in the measurements.

values of average queuing delays then outgoing segments. This result can be interpreted by a reasoning that the amount of data downloaded from the Internet to an organization is usually higher, than the amount downloaded from the servers of the organization by clients situated elsewhere in the Internet. Looking at the spatial arrangement of the state variables, one can see that the internal segments that are connections between branching nodes constitute a core which is characterized by the smallest values of the state variables. This is not surprising, since these are network segments in the gigabit backbone.

After the identification of various spatial structure in the data, we can also investigate its temporal variation. For this purpose we plot in Fig. 4 the average and standard deviation of queuing delays in the case of the three subsequent measurements, for all the different network segments and end-to-end paths. It is interesting that a set of network segments and end-to-end paths does not show any temporal change, while big changes are apparent for another set. Also in the case of end-to-end paths with a common source, a single network segment can be identified that is responsible for the origin of the temporal variation. As expected the measurement at UTC:03:30 shows the smallest values in all metrics.

For the further analysis of the results in Fig. 5 we plot

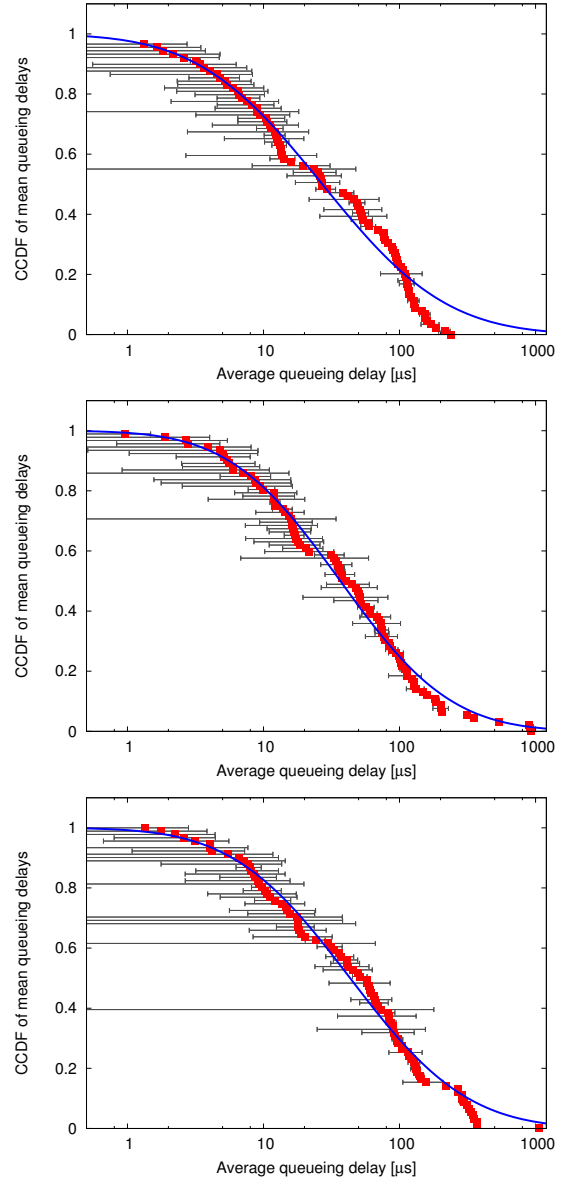


Fig. 5. The complementary cumulative distribution function of the mean queuing delays in the different measurements. On the top UTC:03:30, in the middle UTC:14:05, on the bottom UTC:16:12. The continuous line is a fit assuming log-normal distribution.

the complementary cumulative distribution function of the average queuing delays on the different segments. Since our tomography method ([18], [19]) gives multiple estimations for each network segment, we use error bars to represent the standard deviation of the average queuing delays on different measurement routes. The implication of the very good fit in Fig. 5 in the case of UTC:14:05 is that the average queuing delay of the different segments at the most loaded time of day follows a log-normal distribution

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - m)^2}{2\sigma^2}\right], \quad (1)$$

with parameters  $\sigma \approx 1.42$ , and  $m \approx \ln(37.8\mu s)$ .

## V. CONCLUSION

This paper presented large-scale measurements of queuing delay distributions in a part of the European Internet, conducted via the ETOMIC measurement infrastructure, using special active probing techniques and the methods of network tomography. This very precise and fully synchronized infrastructure meets the requirements needed to perform large-scale unicast tomography measurements, and can be viewed as the prototype of network testbeds, that will be able to operate in the uncooperative Internet of the future. As the main result we presented data of the averages and standard deviations of queuing delay, and identified various spatial and temporal structures in them. We found that the average queuing delay of network segments spans three orders of magnitude, and its distribution function closely follows a log-normal distribution in the case of the most loaded time of day.

## VI. ACKNOWLEDGEMENTS

The authors thank the partial support of the National Science Foundation (OTKA T37903), the National Office for Research and Technology (NKFP 02/032/2004 and NAP 2005/KCKHA005) and the EU IST FET Complexity EVERGROW Integrated Project.

## REFERENCES

- [1] Active Measurement Project, <http://amp.nlanr.net/>.
- [2] M. Coates and R. Nowak. Network tomography for internal delay estimation. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*. Salt Lake City, May 2001.
- [3] N. Duffield, J. Horowitz, F. L. Presti, and D. Towsley. *Network Delay Tomography from End-to-end Unicast Measurements*. Springer Verlag - Lecture Notes in Computer Science, Berlin, 2001.
- [4] M. Coates and R. Nowak. Network loss inference using unicast end-to-end measurement. In *Proc. ITC Conf. IP Traffic, Modelling and Management*. Monterey, CA, September 2000.
- [5] N. Duffield, F. L. Presti, V. Paxson, and D. Towsley. Inferring link loss using striped unicast probes. In *Proc. of IEEE Infocom 2001*. Anchorage, AK, April 2001.
- [6] R. Cáceres, N. G. Duffield, S. B. Moon, and D. Towsley. Inference of internal loss rates in the mbone. In *Proc. of IEEE/ISOC Global Internet 1999*. December 1999.
- [7] R. Cáceres, N. G. Duffield, J. Horowitz, D. Towsley, and T. Bu. Multicast-based inference of network-internal characteristics: Accuracy of packet loss estimation. In *Proc. of IEEE Infocom 1999*. New York, March 1999.
- [8] N. G. Duffield, F. L. Presti. Multicast inference of packet delay variance at interior network links. In *Proc. of IEEE Infocom 2000*. Tel Aviv, Israel, March 2000.
- [9] F. L. Presti, N. G. Duffield, J. Horowitz, and D. Towsley. Multicast-based inference of network-internal delay distributions. *ACM/IEEE Transactions on Networking*, December 2002.
- [10] T. Bu, N. Duffield, F. L. Presti, and D. Towsley. Network tomography on general topologies. In *Proc. of ACM Sigmetrics 2002*. Marina del Rey, CA, 2002.
- [11] M. Rabbat, R. Nowak, and M. J. Coates. Multiple source, multiple destination network tomography. *IEEE/ACM Transactions on Networking*, December 2002.
- [12] EVERGROW homepage, <http://www.evergrow.org/>.
- [13] ETOMIC homepage, <http://www.etomic.org/>.
- [14] D. Morató, E. Magaña, M. Izal, J. Aracil, F. Naranjo, F. Astiz, U. Alonso, I. Csabai, P. Hąga, G. Simon, J. Stéger, G. Vattay. The European traffic observatory measurement infrastructure (etomic): A testbed for universal active and passive measurements. In *Proc. of Tridentcom 2005*, p283-289, Best Testbed Award. Trento, Italy, February 23 – 25, 2005.
- [15] E. Magaña, D. Morato, M. Izal, J. Aracil, F. Naranjo, F. Astiz, U. Alonso, I. Csabai, P. Hąga, G. Simon, J. Stéger, G. Vattay. The European Traffic Observatory Measurement Infrastructure (ETOMIC). In *Proc. of IPOM 2004*, p165-169. Beijing, China, October 11 – 13, 2004.
- [16] ENDACE homepage, <http://www.endace.com/>.
- [17] GÉANT homepage, <http://www.geant.net/>.
- [18] G. Simon, P. Hąga, G. Vattay, and I. Csabai. A flexible tomography approach for queuing delay distribution inference in communication networks. In *Proc. of IPS-MoMe 2005*. Warsaw, Poland, March 14-15, pages 60-69, 2005.
- [19] G. Simon, J. Stéger, P. Hąga, I. Csabai and G. Vattay. Measuring the Dynamical State of the Internet Large Scale Network Tomography via the ETOMIC Infrastructure. *ComplexUs Journal*, to appear, 2006.