Large Scale Network Tomography in Practice: Queueing Delay Distribution Inference in the ETOMIC Testbed

Péter Mátray*, Gábor Simon, József Stéger, István Csabai and Gábor Vattay
Department of Physics of Complex Systems, Eötvös University, Budapest, Hungary
Email: matray@complex.elte.hu

***Abstract*— This poster proposal presents operational experience of large-scale unicast network tomography, that samples a part of the European Internet. We describe in detail the ETOMIC measurement platform that was used to conduct the experiments, and its potential in future scaled-up measurements. Our main results are maps showing various spatial and temporal structure in the characteristics of the inferred queueing delay distributions. At the most loaded time of day we find that the distribution of average queueing delays among the different path segments follows closely a log-normal distribution.**

I. Introduction

The measuring of the static and dynamical state of the Internet is an important and inevitable task for predicting the quality of various services and applications. From the point of view of the current, most widely used 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. An attractive way to measure packet loss-rate and delay over the Internet, is by means of active probing. This technique is flexible and has a wide range of applicability. In the past years several measurement platforms have been developed for conducting active measurements over the Internet, however these platforms can provide only end-to-end information between the participating nodes, or 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 unicast network tomography, which is a special class of active-probing techniques, that is able to resolve the end-to-end delay statistics [1], [2] and packet loss rates [3], [4] 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. By increasing the number of sources and receivers involved in a measurement, in the limiting case the tomography approach resolves completely

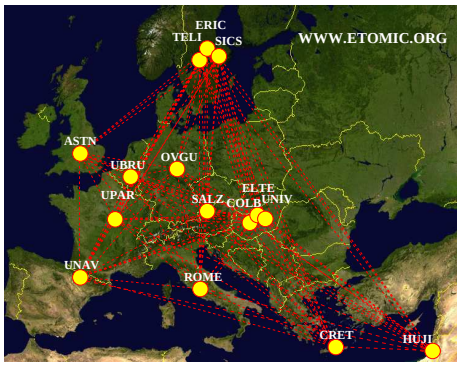

Fig. 1. The geographical locations of the deployed ETOMIC measurement nodes. The abbreviations of the nodes together with additional information are given in Table I.

the network on true links, while end-to-end measurements do not share this property. 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, is the key to resolve the internal characteristics from the end-to-end measurements.

In this proposal we are concerned with large-scale inference of queueing delay distributions by performing extensive unicast network tomography measurements. 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.

II. The ETOMIC Measurement Platform

To perform unicast queueing delay tomography in a real network environment poses several challenges. First in order to

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

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 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 we are developing a state of the art high-precision, synchronized measurement platform, the Evergrow Traffic Observatory Measurement InfrastruCture (ETOMIC)[5]. 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 15 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 [6]. An ETOMIC measurement node is basically a standard PC, 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 measurements. 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. In addition to the above, all the ETOMIC measuring nodes are connected to a GPS, that provides a reference signal directly to the DAG card, thus global synchronization of the nodes can be achieved. Lab experiments performed in advance of the deployment of the nodes revealed an accuracy of one-way delay measurements to be ≈ 0.5 μs, which is mainly limited by the performance of the GPS receivers. By performing extensive traceroute measurements between the available ETOMIC nodes we are able to determine the

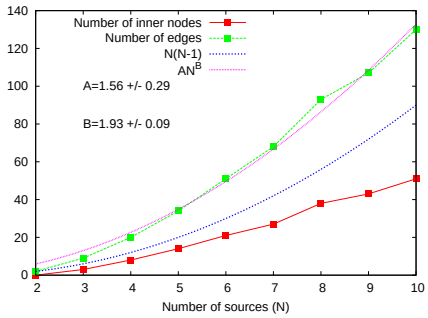

Fig. 2. The scaling of the number of network segments (edges) and branching (inner) nodes with the number ETOMIC source nodes.

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-Gigabit European academic network 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 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. 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 an example measurement performed at a relatively loaded time of day (14:05). The measurement involved 9 ETOMIC nodes, and a connectivity graph consisting of 38 branching nodes and 93 network segments. Each of the source nodes periodically 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. 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. These data sets comprised the input to the tomography method, that yielded the queueing delay distributions resolved 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

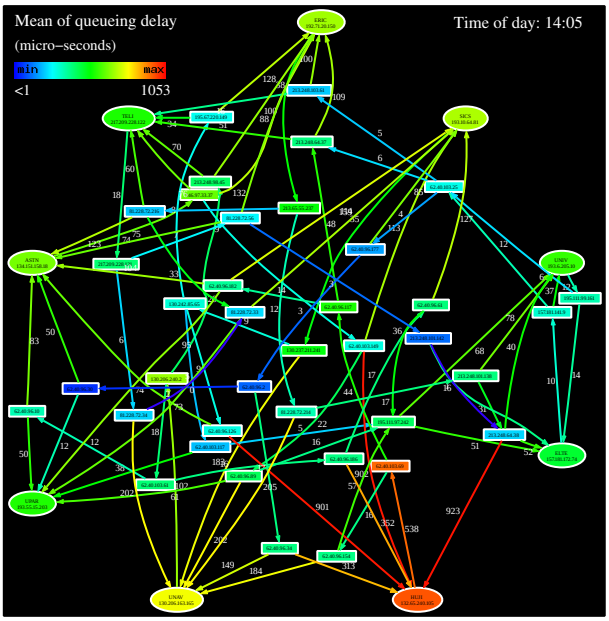

Fig. 3. The connectivity graph colored and labelled by the mean queueing delay, given in units of μs, for each network segment. The ellipse shaped nodes on the edge are ETOMIC measurement nodes with abbreviations given in Table I, while box-shaped nodes in the interior of the graph are branching nodes. The arrows indicate the direction of probe packet flow on a given network segment.

on the quantization of the measured end-to-end delays into 10μs 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 [7] that also provides performance evaluations in ns-2 simulations and controlled lab experiments.

For better visualization of the results we extracted the mean of the queueing delay distributions (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. 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 ($\approx 0.5\mu$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 a 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 queueing delays. Looking at the spatial arrangement of the data, 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.

For the further analysis of the results in Fig. 4 we plot the complementary cumulative distribution function of the average queueing delays on the different segments. The implication of the very good fit is that the average queueing delay of the

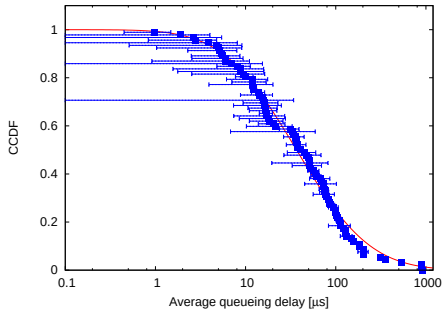

Fig. 4. The complementary cumulative distribution function of the mean queueing delays. The continuous line is a fit assuming log-normal distribution.

different segments follows a log-normal distribution

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

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

IV. Summary

This poster proposal presented large-scale measurement of queueing delay distributions in a part of the European Internet. The measurements were conducted via the ETOMIC testbed, 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. We found that the average queueing delay of network segments spans three orders of magnitude, and its distribution among the different segments closely follows a log-normal distribution in the case of the most loaded time of day.

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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] ETOMIC homepage, http://www.etomic.org/.

[6] D. Morató, E. Magaña, M. Izal, J. Aracil, F. Naranjo, F. Astiz, U. Alonso, I. Csabai, P. Hága, 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*. Trento, Italy, February 23, 2005.

[7] G. Simon, P. Hága, G. Vattay, and I. Csabai. A flexible tomography approach for queueing delay distribution inference in communication networks. In *Proc. of IPS-MoMe 2005*. Warsaw, Poland, March 14-15, pages 60-69, 2005.