Large Scale Internet Queueing Delay Tomography

Yuval Shavitt, Eran Shir
Dept. of Electrical Engineering,
Tel-Aviv University, Tel-Aviv, Israel

Jozsef Steger, Gabor Simon, Gabor Vattay, Istvan Csabai
Dept. of Physics of Complex Systems,
Eotvos Lorand University, Budapest, Hungary

***Abstract*— Queuing delay tomography of the Internet is mostly a theoretical research topic, and measurements were mainly performed to prove the validity of a certain measurement methods. We propose a large scale Internet tomography survey to map the queueing delay in the European networks in great details, and the rest of the world to a lesser extent. The measurements will be based on the ETOMIC high accuracy packet capturing infrastructure and on DIMES vast distributed agent community. We present the rational behind the effort, the new technical tools developed to enable it, and some results from initial trials.**

I. Introduction

The central role the Internet has in today's web of life calls for understanding its behavior and engineering. A key aspect of this understanding is to track the way congestion appears and evolve in the Internet, both in time and space. Measuring congestion directly at every link is a task possible only by a collaborative effort of all network administrators, a situation not possible in the current atmosphere where even the topological structure of networks is regarded by some network administrators secretive.

To be able to better understand congestion in the Internet several techniques have been proposed, which can be classified into two main groups: active probing, when probe packets are injected into the network and the received probe stream is analyzed; and passive monitoring where the traffic flowing through a network element is sampled for analysis. The former technique has a wide range of applicability, however its extensive usage may impose unnecessary load on the network biasing the system. On the other hand passive monitoring does not impose any load, but the researcher must have an access to the network element under study, which is generally not the case. In the past years several measurement platforms have been developed for conducting active measurements over the Internet (e.g. Surveyor, Felix, AMP), however these platforms can provide only end-to-end information between the participating nodes, which can not be resolved on the parts in between. Many of these active probing tools rely on extra cooperation of the routers along the path of a probing stream, which in the future is foreseen to be unavailable due to the decentralized and heterogenous administration.

Network tomography is a means to bypass this problem. It is a special class of active-probing measurement techniques that is able to resolve the end-to-end delay statistics [1], [2] and/or packet loss rates of the internal segments of the paths. Tomography measurement is conducted from a single source to a set of receivers to yield the delay statistics and loss-rates on each segment of the underlying tree spanned by the source and receiver nodes, and the branching nodes (routers where the paths of probes destined for different receivers diverge). By increasing the number of sources and receivers involved in a tomography measurement, the portion of the network for which state information can be resolved grows dramatically.

In this paper, we present a wide scale effort to perform such a tomography survey in Europe and to less extend in the rest of the world. In the next section we present the ETOMIC and DIMES measurement systems, which were are used in our experiment. In Section III the cooperative tomography measurement experiment is described. Next in Section 3 we proceed with the presentation of our initial results, while in the last section we discuss the future of this effort.

II. The ETOMIC System and the DIMES Platform

Much of the early tomography research concentrated on using muticast, which in unfortunately not widely enabled in the Internet. Thus, in this paper we use unicast network tomography, the main idea behind which is to use back-to-back packet pairs, each pair destined to different receivers. As the packet pairs traverse the common portion of the path they experience the same network conditions bringing correlation into the time-series of the end-to-end characteristics. This inherent correlation property is the key to resolve the characteristics of the links from the source of the probes to the branching router and from the branching router to the destination agents. The delay experienced by a packet over an Internet path sums up from two components, a constant propagation delay characteristic of the media and a time-varying component arising from queueing process in the buffers of routers. In this paper we are concerned with the inference of queueing delay distributions. For the propagation delay inference see [3].

It is a challenge to perform unicast queueing delay tomography in a real network environment. 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. In our case it is enough to have synchronized clock references in the receiving agents because the probe stream injected by the sender can be used to calibrate its clock-rate with high precision. The measuring infrastructure, additionally, has to be very precise to be able to resolve the microsecond-scale queueing delay components associated with high-bandwidth (multi-Gigabit) links. To achieve sub-microsecond precision, a unique hardware solution is inevitable. We have developed

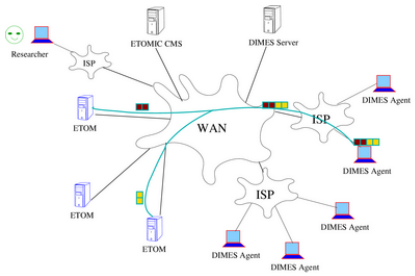


Fig. 1. A unicast based experiment with packet train correlation.

a state of the art high-precision, synchronized measurement platform, the EVERGROW Traffic Observatory Measurement InfrastruCture (ETOMIC). The ETOMIC measuring nodes deployed at different European countries are managed through a central management system that is accessible to researchers through a web-based graphical user interface [4]–[6] providing flexible means to perform various measurement types. The ETOMIC measurement node is basically a standard PC hardware equipped with an Endace DAG 3.6GE card as the network monitoring interface, which is designed for precise active and passive measurements. DAG cards provide very accurate time-stamping of packets captured or transmitted by them, with a time-resolution of 60ns. The measurement nodes and the independent clocks residing in the DAG cards are synchronized by GPS (Garmin GPS 35 HVS) providing a PPS (pulse per second) reference signal. The overall precision of a one-way delay measurement within ETOMIC nodes was found to be 0.5 μs.

DIMES is a distributed Internet measurement project [7] aimed at studying the structure and topology of the Internet with the help of a volunteer community. Over 7500 agents were downloaded by about 4000 volunteers in about 85 countries world wide. The DIMES agents are controlled from a central experiment management system which instruct them to perform measurements to various locations. Data collected by the agent are frequently uploaded to the DIMES databases. The agent measurement traffic is bounded to be below 1Kbps to minimize the load on the hosting machine and on the network around it. Initially two measurement types were implemented `ping` and `traceroute`, which are common tools to infer information about the network connectivity, delay, and loss. For the purpose of this experiment, we added a new module to the agents that enables the emission of packet trains, which are used here to conduct tomography measurement using these two platforms, as described in the next section.

III. The measurement

A new packet train measurement module enables the submission of several trains of packets from a DIMES agent, which in general is not clock synchronized to a group ETOMIC boxes which has a high precision time stamping accuracy and which are synchronized to each other using the GPS. A typical experiment has the following stages. First, the DIMES agents selected as sources discover the paths to the ETOMIC sinks using traceroute. Using the traceroute results, with the DIMES interface to router database the experiment router level topology is built. Next, the ETOMIC boxes are instructed to capture and timestamp the probing stream. In Fig. 1 a schematic view of the experiment is given. A researcher using the web interface of the ETOMIC central management system selects sinks from the set of free ETOMIC nodes and reserves them for their measurement. The DIMES experiment planner provides the researcher a list of statistically available DIMES agents with their approximate geographical location information in order to be able to plan the tomography measurement. When the experiment reported in this paper was performed we lacked this selection interface and the DIMES agent was chosen and instructed directly through the DIMES server.

DIMES agents connect to the DIMES server with regular frequency, data collected from former measurements are uploaded and new tasks are queried during this conversation. Measurement queries are written in a pseudo scripting language, called PENny. The new measurement module we implemented has now a unique call in PENny. In the test tomography measurement we followed the sample script provided here, with the arguments filled in with the appropriate values.

```
<Penny>
 <Script id="trt_ptrain" ExID="1">
  <Priority>URGENT</Priority>
  TRACEROUTE <IP1>
  TRACEROUTE <IP2>
  TRACEROUTE <IP3>
  PACKETTRAIN <nr> <dt> <pl> <pt> <dp> <IP1> [<IP2> [...]]
 </Script>
</Penny>
```

In the script above `Priority` biases the order of measurements queued in the server. `TRACEROUTE` and `PACKETTRAIN` are the names of the corresponding measurement calls, `<nr>` defines how many batches to send, `<dt>` sets the time elapsed between two consecutive batches. Packet size is defined by the argument `<pl>`, `<pt>` selects the Internet protocol, which could be `UDP` or `ICMP`, `<dp>` sets the destination port and `<IP1> [<IP2> [...]]` is the list of IP addresses to form the pattern of each batch.

The test measurement was conducted on the 12th July 2005 starting at 10:24 AM UTC. In the experiment a single DIMES agent was used running on a desktop computer (ELTE), and six ETOMIC measurement nodes were targeted: Vetenarian University, Budapest (UNIV); Ericsson Research, Stockholm (ERIC); Aston University, Birmingham (ASTN); Navarra University, Pamplona (UNAV); Paris University (UPAR), and Hebrew University, Jerusalem (HUJI). Right after topology discovery 10^4 six packet batches were injected into the network with 0.1 seconds gaps between consecutive batches. The six back-to-back packets in the batch were destined to the different Etomic measurement nodes each sent out in a round robin fashion.

Time and topology data collected in the measurement need to be preprocessed before the tomography resolution techniques can be applied. First the clock-skew of the Dimes agent needs to be compensated for. In case the time series is long

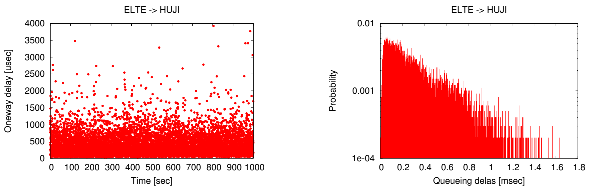
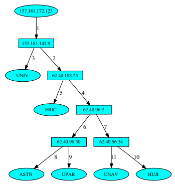


Fig. 2. **The end-to-end queueing delay** Detrended one-way delay time series and its distribution are presented for the link between a DIMES agent (ELTE) and an ETOMIC measurement node (HUJI).



Link	Avg. q.d.(μs)	Error	Std dev.(μs)	Error
1	23.446	2.738	35.044	4.652
2	6.976	3.883	5.052	2.048
3	32.196	2.779	70.856	2.678
4	6.693	4.016	17.467	10.557
5	83.373	5.730	48.426	1.373
6	1.802	0.707	6.059	4.215
7	8.273	1.452	5.441	0.981
8	43.575	0	30.752	0
9	30.866	0	37.188	0
10	309.853	0	294.230	0
11	92.723	0	136.885	0

Fig. 3. **Topology** from the DIMES agent to the six target ETOMIC nodes. The constructed tree contains eleven edges. In the table the links resolved are enumerated. The mean queueing delays and their standard deviations are given along with the corresponding errors.

enough, the measure of the fluctuation of the end-to-end delay series is negligible, and it can be detrended. In our case the clock skew was 19 $\mu s/s$, which we determined by fitting a line to the lower envelope of the one-way delay series. Outliers are dropped from the series (1%). the reconstructed one-way delay series and its distribution are shown on Fig. 2.

Next from the `traceroute` information we determined the spanning tree, which contained five branching nodes and eleven segments that are resolvable by the network tomography, see left of Fig. 3 Raw `traceroute` contained 49 router addresses along the paths to the ETOMIC measurement nodes and 56 links between them.

IV. Analysis of the test-dataset

The determination of the queueing delay distributions were performed by the same technique described in [2] in details.

In Fig. 4 we present two examples of the queueing delay distributions of the 11 resolved edges. The average and the standard deviation of all the queueing delays are shown in the table right of Fig. 3. The characteristic figures span over two orders of magnitude, ranging from average queueing delays of a microsecond scale to the 0.1 millisecond. Segments of the Internet that are described by little delay belong to the inter-academic network, where very fast 1-10 Gigabit/s links are used. It is observable, that the average and the standard deviation of the queueing delay is higher at segments connecting end user nodes to the Internet.

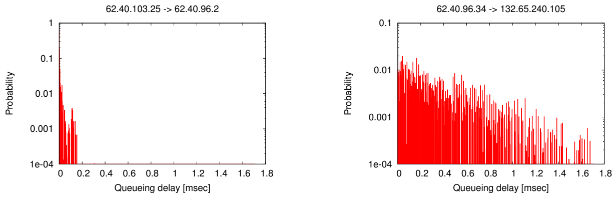


Fig. 4. **Delay distributions** In the measurement eleven edges are resolved between branching points of routers and the agents. Two typical queueing delay distributions are shown in the figure following the link numbering used in Fig. 3: A narrow distribution (4) is typical of in the core network while a wider distribution characterizes access links (10).

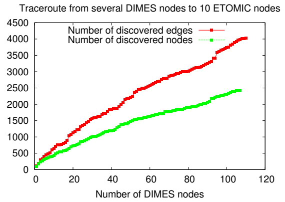


Fig. 5. **Topology statistics, based on DIMES `traceroute` data July 2004.** The number of routers and links discovered by the DIMES agents on their paths to the ETOMIC nodes as a function of the number of agents used.

V. Future perspectives

To test the validity of our approach to discover the dynamics of a large portion of the Internet, a group of DIMES agents were instructed to discover Internet topology between them and ten ETOMIC nodes. Agents having done `traceroute` experiments were ordered according to their first discovery of a new internal router node along the path to an ETOMIC node. Counting the number of known router nodes in this sequence we can reconstruct how many nodes can be discovered by a number of DIMES agents. The same analysis was carried out with link information. As it can be seen in Fig. 5 both the number of routers and the number of internal links increase fast with the number of agents used: only a few agents discover thousands of links.

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] G. Simon, P. Haga, G. Vattay and I. Csabai. A Flexible Tomography Approach for Queueing Delay Distribution Inference in Communication Networks *Proc. of IPS-MoMe, p60-69.*, Warshaw, Poland, March 2005.

[3] Y. Shavitt, X. Sun, A. Wool and B. Yener. Computing the Unmeasured: An Algebraic Approach to Internet Mapping, *Jsac, p67-78, Vol.22, 2004*

[4] Etomic homepage *http://www.etomic.org*

[5] E. Magana, D. Morato, M. Izal, J. Aracil, F. Naranjo, F. Astiz, U. Alonzo; I. Csabai, P. Haga, G. Simon, J. Steger, G. Vattay. The European Traffic Observatory Measurement Infrastructure (ETOMIC) *IPOM 2004, Beijing, China*, October 11-13, 2004

[6] D. Morato, E. Magana, M. Izal, J. Aracil, F. Naranjo, F. Astiz, U. Alonso, I. Csabai, P. Haga, G. Simon, J. Steger and G. Vattay. ETOMIC: A testbed for universal active and passive measurements *TRIDENTCOM 2005, Trento, Italy*, February 23-25, 2005

[7] Dimes homepage *http://www.netdimes.org*