View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Infoscience - École polytechnique fédérale de Lausanne

The Performance of Measurement-Based Overlay Networks

Daniel Bauer, Sean Rooney, Paolo Scotton, Sonja Buchegger, and Ilias Iliadis

<dnb,sro,psc,sob,ili>@zurich.ibm.com
IBM Research, Zurich Laboratory
Säumerstrasse 4
8803 Rüschlikon, Switzerland

Abstract. The literature contains propositions for the use of overlay networks to supplement the normal IP routing functions with higher-level information in order to improve aspects of network behavior. We consider the use of such an overlay to optimize the end-to-end behavior of some special traffic flows. Measurements are used both to construct the virtual links of the overlay and to establish the link costs for use in a link-state routing protocol. The overlay attempts to forward certain packets over the least congested rather than the shortest path. We present simulation results showing that contrary to common belief overlay networks are not always beneficial and can be detrimental.

Keywords: Overlay network, QoS Routing, Link measurement

1 Introduction

Quality of Service (QoS) in large networks is achievable through the presence of control logic for allocating resources at network nodes coupled with inter-router coordination protocols. The various approaches — ATM, DiffServ, IntServ — differ in the trade-off between the precision with which the behavior of flows can be specified and the cost of the additional control logic. However, none of the approaches are widely used in the public Internet. Increased network capacity has meant that the benefits of resource guarantees are reduced and consequently outweighed by the management overhead. Moreover, for HTTP-type request/response traffic this is unlikely to change as the majority of the delay incurred is in the servers [1] rather than in the network, so network guarantees for such flows are of marginal importance.

Applications in which the timeliness of the arrival of data is important, such as continuous media streams, distributed games and sensor applications, would benefit from resource guarantees. Whereas the fraction of Internet traffic that such applications constitute may increase, it is unlikely that this increase will be sufficient to force internet service providers (ISPs) to instrument flow or aggregated flow guarantees. Moreover, it would involve the difficult coordination of policy between the border gateways of autonomous systems of different ISPs.

In an overlay network higher-layer control and forwarding functions are built on top of those of the underlying network in order to achieve a special behavior for certain traffic classes. Nodes of such a network may be entities other than IP routers, and typically these networks have a topology that is distinct from that of the underlying physical network. Nodes in the overlay network use the IP network to carry both their data and control information but have their own forwarding tables as a basis for routing decisions. Examples of overlays are the Gnutella file-sharing network and the Mbone multicast network.

Our approach is to treat traffic requiring guarantees as the special case rather than the common one. This special traffic is forwarded between network servers with hardware-based packet forwarding across a dedicated overlay. We call these servers *booster boxes* [2]. The routing logic between the booster boxes uses dynamic measurements and prediction to determine the least congested path over the overlay. Traffic that is carried over the booster overlay network is called overlay traffic.

While it is trivial to describe simple idealized scenarios in which overlays bring a gain, the more pertinent question is whether and under what circumstances measurement-based overlay networks are beneficial in realistic networks. The focus of this paper is on the applicability and performance of a measurement-based overlay network.

The remainder of the paper is organized as follows. After a review of related work in Section 2, we outline the general architecture of such an overlay network of booster boxes in Section 3. In Section 4 we describe our detailed simulation of its behavior in diverse scenarios. The simulation results and discussion can be found in Section 5, followed by the conclusions in Section 6.

2 Related Work

The resilient overlay network (RON) architecture [3] addresses the problem that network outages and congestion result in poor performance of IP routing and long recovery time due to slow convergence of Internet routing protocols such as BGP. RON uses active probing and passive monitoring in a fully meshed topology in order to detect network problems within seconds. Experimental results from a small real-world deployment of a RON have been obtained demonstrating fast recovery from failure and improved latency and loss rates. Note that the authors do not claim that their results are representative of anything other than their deployment and no general results for different topologies, increased RON traffic, etc., have been published.

The Detour [4] framework pointed out several routing inefficiencies in the Internet and mainly attributed them to poor routing metrics, restrictive routing policies, manual load balancing, and single-path routing. By comparing actual routes with measurement traces between hosts, Savage et al. found that in almost every case there would have been a better alternative path. They envision an overlay network based on IP tunnels to prototype new routing algorithms on top of the existing Internet; however, they concede that measurement-based adaptive routing can lead to instability and, to the best of our knowledge, no evaluation of the overlay performance has been published.

A different application of overlay networks is content-based navigation in peer-to-peer networks. The goal of content-addressable networks such as Chord, CAN [5], Tapestry, and Pastry [6] is efficient, fault-tolerant routing, object location and load-balancing within a self-organizing overlay network. These approaches provide a scalable fault-tolerant distributed hash table enabling item location within a small number of hops. These overlay networks exploit network proximity in the underlying Internet. Most use a separate address space to reflect network proximity.

Although these overlay networks have been shown to work in some specific cases, no extensive simulations or practical measurements on a wide range of topologies have been carried out.

3 Architectural Overview

In this section we briefly outline the overlay architecture which we evaluate by simulation. The overlay network consists of a set of booster boxes interconnected by IP tunnels. Packets are forwarded across virtual links, i.e. the IP tunnels, using normal IP routing. The IP routers are not modified and are entirely unaware of the existence of the overlay network.

Booster boxes that are directly connected across a virtual link are called peers. Note that a single virtual link may correspond to multiple IP paths. Booster boxes peer with other booster boxes with which they are likely to have good connectivity. This is determined using pathchar [7] and/or packet tailgating [8]. Pathchar provides more information than packet tailgating about the entire path but has more restrictive assumptions. Both techniques are known to fail beyond a certain threshold number of hops, because of error amplification. We therefore restrict the hop count of the virtual links.

Although the establishment of a virtual link between two booster boxes is asymmetrical, both sides must agree to the peering. We use this, together with the fact that boosters box have good knowledge of the links to which they are directly attached, to determine the accuracy of the link measurement.

The techniques for determining link characteristics require the transmission of a large number of packets and accordingly take a significant amount of time to determine a result. They are adequate for the construction of the overlay network but not for the transient state links measurements used to make forwarding decisions. Booster boxes, on the other hand, measure the current latency and loss probability of the virtual links by periodically exchanging network probes with their peers. This is similar to the Network Weather Service described in [9].

Booster boxes maintain the overlay network forwarding tables using a link-state routing protocol. If the link state on a booster box changes significantly, the forwarding tables are recomputed. Packets are forwarded between booster boxes using classical encapsulation techniques.

4 Overlay-Network Simulation

The simulation process contains the following steps. First, we generate a physical network topology using the Brite [10] topology generator. The result is a graph consisting of nodes that represent autonomous systems (ASs) and of edges that represent network links with certain capacities and delays. In a second step, we populate the network with applications with four sources sending a constant stream of packets to a single sink using UDP as the transport protocol; this corresponds to a sensor-type application. To generate "background" traffic and thus congestion we add several TCP sources. The result of this step is a TCL script that is fed into the NS-2 network simulator [11]. Then, we create an overlay network by adding booster boxes to the network topology. Finally, a small fraction of the applications are reconfigured such that they send traffic over the overlay network. The result of this last step is another TCL script, executable by the NS-2 network simulator.

The traffic generated by the applications is analyzed. In particular, we are interested in the average packet-drop ratio, i.e. the ratio of dropped packets to the total number of packets sent. For a given topology, we obtain three results. The first is used as reference and is obtained when no overlay network is present. With an overlay network present, we obtain a second drop ratio of those applications that do not use the overlay, and a third result which is the drop ratio of those applications that use the overlay network. Figure 1 shows the four steps involved and the resulting two experiments.

Physical Network Topology. We generate random topologies using the Waxman model with parameters set to $\alpha = 0.9$ and $\beta = 0.2$. A high value for α was chosen to prioritize local connections. The ratio of nodes to links is 1:2. The link capacity varies randomly from 1 to 4 Mb/s and link propagation delay in the range of 1 to 10 ms. We consider network sizes of 100, 200, and 400 ASs. These are rather small compared to the Internet but we are constrained by the performance of the NS-2 tool. It would perhaps be more realistic to use the power law [12] rule of Internet ASs, but our networks are too small for this to be feasible.

We assume the physical topology to be invariant during a run of the simulation, i.e. nodes and links do not fail and therefore no dynamic routing protocol is needed. Given the fact that routing convergence in the Internet using BGP

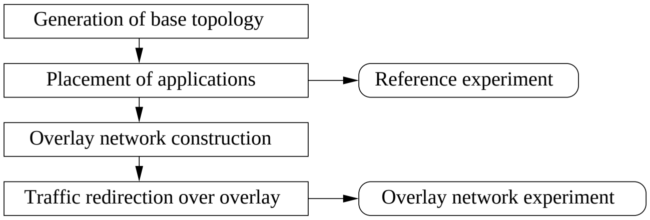
Generation of base topology
Placement of applications
Reference experiment
Overlay network construction
Traffic redirection over overlay
Overlay network experiment

Fig. 1. Steps involved in a single simulation run

is rather slow [13] compared with the convergence time of the link-state routing protocol used in the overlay network, it would be expected that the overlay would react to failure more quickly than the physical network does.

Overlay Network Topology. The overlay network is constructed by adding booster boxes to the ASs with the highest degree of connectivity. This is to ensure that booster boxes are mainly placed in transit ASs. Booster boxes construct virtual links to the four closest neighboring booster boxes. Closeness is measured in terms of hop count, and the path capacity is used as a tie-breaker in the case of equal hop counts. In our simulation model, an AS is equipped with a booster box by creating an additional node of type "booster box" and connecting it to a single AS node using a high-capacity link of 100 Mb/s and a latency of 100 μs. Inter-AS communication has much longer delay and is more susceptible to packet loss than AS-booster box communication. We carry out our simulation across a small range of booster/AS ratios: 1:10, 1:5, and 1:2.

Traffic Characterization. Each application consists of four sources sending a constant bit-rate stream of UDP packets to a single sink. The bit rates of the applications follow a normal distribution with mean of 250 kb/s and standard deviation of 50 kb/s. The packet size is 576 bytes for all applications, as this is the predominant data packet size in the Internet. The background TCP traffic has exponentially distributed burst length. The idle time is also exponentially distributed with the same mean as the burst length. To reflect the diurnal nature of Internet traffic we introduce periods where the number of background TCP traffic sources is double. In different experiments, we vary the burst length of the background traffic such that their mean is either 1 ms, 10 ms, 100 ms, or 1000 ms. Figure 2 shows the effect of the diurnal traffic model over a link that carries 10 TCP streams with burst length 1 ms. The application and background traffic sources and sinks are allocated at the edge of the network. This is done by randomly distributing the sources and sinks among the 60% of the ASs that have the lowest connectivity. The number of application sources is 0.4 times the number of nodes, for background traffic this factor is 5. We do not attempt a realistic characterization of traffic produced by an AS, but simply try to ensure that congestion occurs at arbitrary times and for different periods.

Ratio of Overlay Traffic. A small fraction of traffic produced by the applications is sent over the overlay network. In our experiments, this ratio ranged from 5% to 12%. As the applications themselves only produce 5% of all the traffic, the ratio of overlay traffic to the total traffic is in the range of 0.3% to 0.6%.

Measurement of the Dynamic Metrics. In the experiments the cost of an overlay link is a linear function of the TCP smoothed RTT (SRTT) as measured between each booster box and its peers. Using the SRTT prevents the link cost from oscillating wildly. As the RTT is not updated using retransmitted packets [14] when TCP times out, and when therefore potentially a packet has been lost, we set the cost of the link to a much higher value than any observed RTT; in this way more weight is attached to loss than delay. We send a single 50-byte

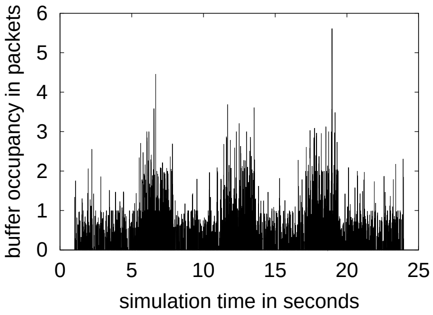

Fig. 2. Buffer occupancy per time for the diurnal traffic model

probe every 500 ms between peered booster boxes. For reasons of simplicity, the simulation does not use a predictive model for attempting to identify future values of the SRTT such as those described in [9].

Frequency of Virtual Routing Exchanges. The booster-box routing agents check every second their link-state values. They exchange link state updates with their peers if the costs associated with their links have changed.

Simulation Scenarios. We divided the experiments into three sets:

- *Light congestion*: 500 background traffic sources using a normally distributed bit rate with a mean of 250 kb/s and a standard deviation of 50 kb/s and 40 application sources are used. Only the data generated by four of the application sources is routed through the overlay network. In this scenario we have a maximum of 4% packet loss in the reference experiments.
- *Heavy congestion*: the mean rate and the standard deviation of the background traffic is doubled to 500 kb/s and 100 kb/s respectively. Here the packets losses in the reference experiment reach 36%. While such high losses are unusual, they have been observed on backbone routers [15].
- *Low overlay usage*: the same as *light congestion* except that for the applications using the overlay network, only the data generated by two of its sources is routed through the overlay network.

For each set we run ten experiments varying the number of booster boxes in the network and the burst length of the background traffic. Each experiment simulates 24 seconds of network operation. A different topology is generated for each experiment.

5 Results

We do not observe significant variation in the results owing to the different topologies sizes — perhaps due to the fact that they are only scale orders of difference — and therefore we only present those for 100 nodes. Tables 5 shows the experimental results for the three sets. The tables contain the following information for each booster box ratio and burst length couple:

- The column labeled "Ovl vs Ref" shows the percentage of the experiments in which the loss ratio of the traffic using the overlay network is smaller than that of the traffic in the reference experiments.
- The column labeled "Norm vs Ref" shows the percentage of the experiments in which the loss ratio of the traffic not using the overlay network is smaller than that of the traffic in the reference experiments.
- The three last columns show the average loss ratio and, in parenthesis, the standard deviation of all the experiments for traffic using the overlay network; traffic not using the overlay network, and traffic in the reference experiments.

We do not think the mean of the packet drop ratio is representative due to the high variance of the results. More interesting is the number of experiments which show a benefit. For example, the first row of the table reports an average drop rate of 8.5% for the overlay traffic, but only a drop rate of 3.6% for the reference traffic. On the other hand, in 40% of the experiments, the overlay traffic had a lower drop rate.

An overlay network is *beneficial* when the overlay traffic behaves better than the reference traffic and the non-overlay traffic behaves no worse than the reference traffic. An overlay network is *partially beneficial* when the overlay traffic behaves better than the reference traffic but the non-overlay traffic behaves worse than the reference traffic. An overlay network is *detrimental* when the overlay traffic behaves worse than the reference traffic.

All three cases are observed in the results. Our belief is that the detrimental behavior is due to an aggregation effect, in which flows that would have taken different paths over the physical network are forced to take the same one owing to lack of an alternate path in the overlay network. This causes unnecessary congestion and is strongly dependent on both physical and overlay network topologies.

We suppose that the additional traffic needed for overlay routing and measurement is the cause of the partially beneficial results. The traffic not using the overlay network is effected by this overhead without deriving any benefits from it. This effect is worsened by the fact that exchanges of link-state advertisements occur most often in the case of congestion.

The beneficial case is when the overlay succeeds in recognizing and routing around congestion without disrupting other traffic. In some cases both the traffic using the overlay network and that not using it benefit from the overlay, simply because they are routed through different paths.

In general, the highest benefits from the overlay network are observed in the sets of experiments with heavy congestion and low overlay usage. In these two cases the benefit seems to increase the more booster boxes there are. This is not true for the light congestion case. Another remark is that the benefit does not seem to be radically effected by the burstiness of the background traffic. We expected to see a significant difference as the burst length increases allowing more time for the overlay to detect the congestion and react. That this was not so could be due to an artifact of the relative durations of the congestion burst and measurement times. However, it is very difficult to characterize the circumstances in which the overlay is beneficial, as we observe a high variance in the experimental results. This fact is depicted in Fig. 3 for the three simulation scenarios. It compares the drop ratios of the overlay network traffic with those of the reference traffic. A "+1" indicates a better result for the overlay case, a "0" the case where both are equal within ± 5%, and a "-1" the case where the reference traffic had a lower drop ratio. A clear tendency can only be observed in the heavy congestion and to a lesser extend the low overlay usage scenarios

Table 1. Simulation results

light congestion						
#BBoxes	**Burst**	**Ovl vs Ref**	**Norm vs Ref**	**Drops Ovl**	**Drops Norm**	**Drops Ref**
10	1	40	40	8.5 (6.7)	3.1 (0.7)	3.6 (1.8)
10	10	30	60	8.1 (6.9)	2.5 (0.7)	3.1 (1.9)
10	100	50	70	7.8 (6.9)	2.2 (0.8)	2.8 (1.9)
10	1000	50	60	6.4 (6.2)	1.9 (0.7)	2.5 (1.9)
20	1	80	10	3.8 (6.4)	4.3 (1.7)	4.0 (1.7)
20	10	80	20	3.0 (6.1)	3.8 (1.8)	3.4 (1.8)
20	100	90	20	2.4 (5.8)	3.3 (2.0)	3.0 (1.9)
20	1000	90	30	2.5 (6.0)	3.0 (1.8)	2.8 (1.7)
50	1	70	10	3.7 (2.6)	4.0 (1.7)	3.4 (1.6)
50	10	40	0	3.4 (2.4)	3.4 (1.8)	2.9 (1.6)
50	100	50	20	3.0 (2.7)	2.9 (1.9)	2.5 (1.6)
50	1000	50	20	2.6 (2.3)	2.6 (1.8)	2.2 (1.5)
heavy congestion						
#BBoxes	**Burst**	**Ovl vs Ref**	**Norm vs Ref**	**Drops Ovl**	**Drops Norm**	**Drops Ref**
10	1	80	10	20.8 (13.0)	35.9 (5.0)	33.5 (4.8)
10	10	70	0	14.3 (8.4)	20.9 (4.6)	19.2 (4.2)
10	100	50	10	14.0 (8.1)	17.1 (4.5)	15.6 (3.8)
10	1000	80	0	12.0 (7.4)	17.1 (4.9)	15.7 (4.0)
20	1	80	30	26.9 (16.5)	37.4 (7.0)	36.0 (7.3)
20	10	60	20	18.7 (16.0)	22.1 (6.7)	21.2 (7.0)
20	100	60	20	17.2 (14.8)	18.5 (6.4)	17.5 (6.4)
20	1000	60	40	16.3 (14.1)	17.8 (6.1)	17.4 (6.0)
50	1	100	60	13.7 (6.6)	36.1 (6.8)	35.3 (5.6)
50	10	100	40	5.7 (3.4)	21.2 (6.2)	20.8 (5.7)
50	100	100	40	5.5 (3.0)	17.8 (5.9)	17.2 (5.4)
50	1000	100	40	5.2 (2.3)	18.0 (5.8)	17.5 (5.3)
low overlay usage						
#BBoxes	**Burst**	**Ovl vs Ref**	**Norm vs Ref**	**Drops Ovl**	**Drops Norm**	**Drops Ref**
10	1	50	20	5.3 (4.0)	3.9 (2.1)	3.9 (2.3)
10	10	40	10	5.6 (4.1)	4.0 (2.4)	3.7 (2.3)
10	100	50	10	4.9 (4.2)	3.5 (2.3)	3.3 (2.2)
10	1000	60	20	3.7 (3.7)	2.9 (2.0)	2.8 (2.0)
20	1	40	20	4.3 (3.8)	3.4 (1.8)	3.2 (1.7)
20	10	40	10	3.9 (3.2)	3.0 (1.9)	2.7 (1.8)
20	100	40	10	3.7 (4.0)	2.5 (1.9)	2.2 (1.7)
20	1000	40	20	3.1 (3.5)	2.2 (1.8)	1.9 (1.7)
50	1	90	0	0.9 (1.0)	4.4 (1.6)	3.4 (1.3)
50	10	90	0	0.7 (0.6)	3.8 (1.7)	2.9 (1.4)
50	100	90	0	0.7 (0.9)	3.0 (1.6)	2.3 (1.4)
50	1000	80	0	0.5 (0.7)	3.0 (1.7)	2.1 (1.4)

when a large number of booster boxes is used. In the other cases, the behavior exhibits no clear trend.

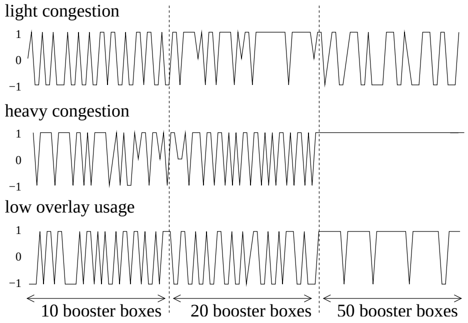

Fig. 3. High Variance in Simulation Results

A distinction can be made between two types of parameters that influence the performance of the overlay: those under the control of the booster-box operators, such as frequency of measurements, routing, or peering strategy, and those not under the control of the booster-box operators, such as network topology or pattern of background traffic. For the first set it is feasible that extensive simulation for precise scenarios would allow useful heuristics to be derived, e.g. never send more than 10% of the total traffic over the overlay. The second are, in general, unknown to the operator. This leads us to conclude that overlays of the type described here need to be reactive, i.e. they need to test the network state and only be activated when they can bring benefit and deactivated otherwise.

While it would be unwise to attach too much importance to simulations that might produce very different results by the simple modification of one parameter, the results show that in the situations tested overlaying can cause significant deterioration of the network. The scenarios may or may not be realistic; however, more simulations and modeling are necessary to better understand the behavior of overlays before they can be deployed.

6 Conclusion

We have outlined an architecture for measurement-based overlay networks that allows certain traffic flows to be privileged over others. We present results from simulation showing how this architecture might behave in the public Internet. We found that while the overlay can be beneficial, it often is detrimental. The circumstances under which the undesired behavior occurs are difficult to characterize and seem very sensitive to small changes in the parameters. As some of these parameters are not under the control of the overlay supervisors, we conclude that such overlays need to be reactive. As a final remark we suggest that proponents of overlay networks need to investigate the effect of their deployment, not only in simple, idealized scenarios, but on the network as a whole.

References

1. Cleary, J., Graham, I., McGregor, T., Pearson, M., Ziedins, I., Curtis, J., Donnelly, S., Martens, J., Martin, S.: High Precision Traffic Measurement. IEEE Communications Magazine **40** (2002) 167–183
2. Bauer, D., Rooney, S., Scotton, P.: Network Infrastructure for Massively Distributed Games. In: NetGames 2002 – First Workshop on Network and System Support for Games, Braunschweig, Germany (2002)
3. Andersen, D.G., Balakrishnan, H., Kaashoek, M.F., Morris, R.: Resilient Overlay Networks. In: Proc. 18th ACM Symposium on Operating Systems Principles, Banff, Canada (2001)
4. Savage, S., Anderson, T., Aggarwal, A., Becker, D., Cardwell, N., Collins, A., Hoffman, E., Snell, J., Vahdat, A., Voelker, G., Zahorjan, J.: Detour: a Case for Informed Internet Routing and Transport. Technical Report TR-98-10-05, University of Washington (1998)
5. Ratnasamy, S., Francis, P., Handley, M., Karp, R., Shenker, S.: A Scalable Content Addressable Network. In: ACM SIGCOMM. (2001) 161–172
6. Rowstron, A., Druschel, P.: Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. In: IFIP/ACM International Conference on Distributed Systems Platforms (Middleware). (2001) 329–350
7. Jacobson, V.: How to Infer the Characteristics of Internet Paths. Presentation to Mathematical Sciences Research Institute (1997) ftp://ftp.ee.lbl.gov/pathchar/msri-talk.pdf.
8. Lai, K., Baker, M.: Measuring link bandwidths using a deterministic model of packet delay. In: ACM SIGCOMM 2000, Stockholm, Sweden. (2000) 283–294
9. Wolski, R., Spring, N., Hayes, J.: The Network Weather Service: A Distributed Resource Performance Forecasting Service for Metacomputing. In: Journal of Future Generation Computing Systems. (1998)
10. Medina, A., Lakhina, A., Matta, I., Byers, J.: BRITE: Universal Topology Generation from a User's Perspective. Technical Report BUCS-TR2001 -003, Boston University (2001)
11. Fall, K., Varadhran, K.: The ns manual (formerly ns Notes and Documentation. http://www.isi.edu/nsnam/ns/ns-documentation.html (2002)
12. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On Power-Law Relationships of the Internet Topology. In: ACM SIGCOMM, Harvard University, Cambridge, US (1999) 251–262
13. Labovitz, C., Ahuja, A., Bose, A., Jahanian, F.: Delayed Internet Routing Convergence. In: ACM SIGCOMM 2000, Stockholm, Sweden. (2000) 175–187
14. Karn, P., Partridge, C.: Improving Round-Trip Time Estimates in Reliable Transport Protocols. Computer Communications Review **17** (1987) 2–7
15. Floyd, S., Paxson, V.: Difficulties in Simulating the Internet. IEEE/ACM Transactions on Networking **9** (2001) 392–403