Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2005-001
MIT-LCS-TR-1009

November 1, 2005

Towards Realizing the Performance and Availability Benefits of a Global Overlay Network

Hariharan Rahul, Mangesh Kasbekar,
Ramesh Sitaraman, Arthur Berger

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Towards Realizing the Performance and Availability Benefits of a Global Overlay Network

Hariharan Rahul MIT CSAIL
Mangesh Kasbekar Akamai Technologies
Ramesh Sitaraman U. Mass., Amherst
Arthur Berger Akamai/MIT CSAIL

Abstract. Prior analyses of the benefits of routing overlays are based on platforms consisting of nodes located primarily in North America, on the academic Internet, and at the edge of the network. This paper is the first global study of the benefits of overlays on the commercial Internet in terms of round trip latencies and availability, using measurements from diverse ISPs distributed across all tiers, and over 1100 locations (77 countries, 630 cities and 6 continents).

Our study shows that while overlays provide some improvements in North America, their benefits are especially significant for paths with Asian endpoints. For example, overlays reduce latencies of a quarter of the paths in Asia by over 50%, while the latencies of over half the paths globally are improved by at least 10%. Furthermore, overlays increase availability in North America and Europe by up to 0.5%, whereas improvements in Asia are significantly larger at 3.25%.

Our study also focuses on practical considerations in constructing overlay routes. We show that a simple algorithm that randomly chooses a small number of alternate redundant paths achieves an availability of over 99.5%. We also propose and evaluate a simple predictive scheme that achieves almost optimal latency using only 2-3 paths. Finally, we demonstrate that this is achievable with surprisingly persistent routing choices.

1 Introduction

Many Internet applications have stringent performance requirements. For example, important classes of traffic such as the web, interactive applications such as remote shells over virtual private networks (VPNs) and emerging technologies such as voice over IP (VoIP) are latency sensitive. Additionally, as described in [3], the current low level of availability of the Internet, as compared to networks such as the public telephone system, makes it unattractive for applications such as medical collaboration and financial transactions. Further, lack of performance, or worse, brief downtime for even a few seconds can result in poor user perception, and significant revenue losses for business that conduct transactions on the Internet [20]. Business trends such as outsourcing and workforce

consolidation [5], as well as stringent requirements for applications such as government communications [10], necessitate that these exacting performance and availability standards are met, not just within a single country or small group of countries, but globally.

Prior work [3, 9] has suggested that routing overlays can enable the Internet to achieve a higher degree of performance and reliability than is obtained by the current Border Gateway Protocol (BGP) routing infrastructure. While these studies have provided us insights into the potential of overlays, they have the following limitations:

- The work has been performed on a platform hosted largely on Internet2, whose capacity and usage patterns, as well as policies and goals, differ significantly from the commercial Internet.
- Overlays used in these studies have a footprint primarily in North America. It is well known that the network interconnectivity and relationships are different in Europe and Asia as compared to the continental United States [8]. As such, the performance of overlays connecting end points in these geographies cannot be predicted from the studies.
- Most of the nodes in these deployments are in edge/stub networks, whereas commercial routing overlays [1] would naturally be largely deployed in core tier-1 and tier-2 networks of the commercial Internet, rather than a large number of disparate edge networks, so as to maximize their reach with a manageable deployment.

This paper is the first study of the performance and availability benefits of routing overlays on the commercial Internet. We use a global subset of the massive Akamai content delivery network (CDN) for data collection. Specifically, we collect measurements from 1100 locations distributed across many different kinds of ISPs in 77 countries, 630 cities, and 6 continents. As will be seen later, the variations of our results across geographies bear out the necessity of an extensive global study.

Our work therefore addresses the limitations of prior work outlined above, by using nodes that receive their connectivity from the commercial Internet. The nodes in our platform are in Europe, Asia and other continents in addition to North America. Moreover, they belong to many diverse ISPs ranging from large tier-1 and tier-2 networks to small edge networks.

We address the problem of picking overlay routes to optimize connections between end users and large servers hosting applications such as web, voice over IP, and games. We investigate the performance benefits for these services, which can be characterized by round trip latency (to the first order), as well as path availability. Applications such as large file downloads whose performance is more accurately characterized by throughput are not addressed in this study.

Our goal is to determine practical strategies for an overlay to achieve latency and availability benefits that are close to the optimal paths using the overlay. We evaluate what improvements are achievable by characterizing the ideal gains that an overlay could achieve if it always chose the best possible path at any instant. We classify these enhancements by the continent pair representing the end user

and the server, to understand the differences in performance and availability as a function of geography.

The ideal benefit computation assumes that the overlay is constructed using instantaneous information about the global Internet. Equivalently, these benefits can also be achieved if the overlay can send the information redundantly over every available direct and indirect path. Neither of these possibilities is achievable in practice, where one must necessarily use information collected at time $t - \tau$ to make routing decisions at time t, where τ is the maximum extent to which information can be stale. Further, for reasons of cost and network friendliness, a practical system can only send data along some small number, κ, of redundant overlay paths simultaneously. While one can imagine many complicated heuristics to trade off κ and τ, we have designed and evaluated simple algorithms to demonstrate that reasonable values for κ and τ achieve levels of performance close to the omniscient overlay with very low cost, across a diversity of networks and geographies. Our analysis also indicates that the number of alternative redundant paths can be reduced significantly in practice, by adaptively choosing the value of κ.

The key contributions of our work are the following:

- It is the first evaluation of an overlay that utilizes data from the commercial Internet. Our study shows that while fewer than 10% of paths in North America see an improvement of over 50% with an overlay, almost a quarter of all Asian paths have their latency reduced by at least half. Similarly, while overlays improve path availability by 0.5% in Europe and North America, routes within Asia see gains larger than 3.25%. As we discuss in Section 7, there are good reasons to believe that the disparities in many of these international groups will persist over time.
- Our study provides useful cross validation for the currently deployed testbeds such as PlanetLab [16] and RON [18]. Our measurements indicate that these deployments provide qualitatively similar data for the commercial Internet in North America, but do not capture the global diversity of network topology, especially in Asia.
- We show that randomly picking a small number of paths ($\kappa = 3$ for Europe and NA, and 5 for Asia) achieves availability gains that approach the optimal. Additionally, we demonstrate that for reasonable values of τ (say, 10 minutes) and $\kappa =)2$ paths, over 90% of paths without endpoints in Asia have latency improvements within 10% of the ideal, whereas paths that originate or end in Asia require 3 paths to reach the same levels of performance.
- We provide strong evidence that overlay choices have a surprisingly high level of persistence over long periods of time (several hours), indicating that relatively infrequent network probing and measurements can provide optimal performance for almost all paths.

1.1 Roadmap

The rest of the paper is organized as follows. Section 2 presents an overview of related work, and outlines the context of our present study. Section 3 for-

mally defines the problem analyzed in this paper, and includes description of our testbed, as well as how measurement data is collected. Sections 4 and 5 provide detailed metrics on the ideal performance and availability gains that can be achieved by overlays in a global context, both for all paths, as well as for paths with particularly poor performance and availability. Section 6 addresses issues in real overlay design, and explores structural and temporal properties of practical overlays for performance and availability. We summarize our study in Section 7, and present some directions for future research.

2 Related Work

There have been many measurement studies of Internet performance and availability. See, for example, the work at the Cooperative Association for Internet Data Analysis (CAIDA) [6], and the National Internet Measurement Infrastructure (NIMI) [14, 15]. Examples of research routing overlay networks are the Resilient Overlay Networks project at MIT [18] and the Detour project at U. of Washington [9]. A commercial routing overlay is offered by Akamai Technologies [1].

Andersen et al. in [4] present the implementation and performance analysis of the routing overlay called Resilient Overlay Networks (RON). They found that 51% of the time, improved latency could be obtained via the overlay. This can be compared to the intra-North-America (NA-NA) row in Table 2 in the present study, where an overlay gives improved latency 63% of the time. They also found a single-hop indirection to be sufficient. The higher improvements obtained in our case are probably largely a consequence of the fact that our overlay testbed has a large core network footprint that can be used for the intermediate hop, whereas the RON deployment is edge centric. Andersen et al. in [3] describe a multi-homed overlay network (MONET) which combines multi-homing at the client location with overlay proxies. Measurements from a six-site deployment of MONET showed that it eliminated almost all network-based failures.

Akella et al. in [2] investigate how well a simpler route-control multi-homing solution compares with an overlay routing solution. Although the focus of the study is different than the present paper, it includes results for a default case of a single-homed site, and the authors find that the overlay routing improves RTTs on average by 25%. The experiment was run using 68 nodes located in 17 cities in the U.S., and can be compared with 110 node, intra-North-America (NA-NA) case herein, where we found that the overall latency improvement to be approximately 21%, although the improvement varies significantly across different continent pairs. See Table 2 below.

Savage et al. [19] using data sets of 20 to 40 nodes found that for roughly 10% of the host pairs, the best alternative has 50% lower latency. We obtained the comparable value of 9% for the case of intra-North America nodes, though significantly disparate results for other continent pairs. Again, please refer to Section 4 and Table 2 for details.

More recently, and in parallel with our evaluation, Gummadi et al. [11] implemented random one-hop source routing on PlanetLab and showed that using up to 4 randomly chosen intermediaries improves the reliability of Internet paths.

The significant difference between our research and these other studies is our comprehensive global analysis of performance and availability gains ideally provided by overlays, as well as a description of algorithms and benefits of a practical, predictive overlay.

3 The experimental setup

We address the problem of optimizing paths between end users and enterprise servers. End users are normally located in small lower tier networks, while enterprise servers are usually hosted in tier one networks. We consider routing overlays comprised of nodes deployed in large tier one networks, which can function as intermediate hops in a path from end users to enterprise servers.

3.1 The measurement platform

The machines of the Akamai CDN are deployed in clusters in several thousand geographic and network locations. A large set of these clusters is located near the edge of the Internet (i.e. close to the end-users in non tier one providers). A smaller set exists near the core ISPs (directly located in tier one providers), which serve a large fraction of end-user traffic. We chose a subset of 1100 clusters from the whole CDN for this experiment, based on geographic and network location diversity, security, and other considerations. These clusters span 6 continents, 77 countries, and 630 cities. Machines in one cluster get their connectivity from a single provider. Approximately 15% of these clusters are located at the core, and the rest are at the edge. The set of edge nodes (clusters) represents end-users (excluding their last-mile connectivity) and the set of core nodes (clusters) is representative of enterprise servers. The intermediate nodes of the overlay (used for the alternate indirect paths) are also limited to the core set. Table 1 shows the geographic distribution of the selected nodes. All the data collection for

Continent (Mnemonic)	Edge set	Core set
Africa (AF)	6	0
Asia (AS)	124	11
Central America (CA)	13	0
Europe (EU)	154	30
North America (NA)	624	110
Oceania (OC)	33	0
South America (SA)	38	0

Table 1. Geographic distribution of the platform

this paper was done in complete isolation from the CDNs usual data collection activity.

3.2 Data collection for performance and availability

Each node in the set of 1100 clusters ran a task that sent ICMP echo requests (pings) of size 64 bytes every 2 minutes to each node in the core set. Each task lasted for 1.5 hours. The output was a set of 4-tuples: `<timestamp, source-id, destination-id, RTT>`. If a packet was lost (no response received within 10 seconds), then a special value is reported as the round-trip latency. Three tasks were run every day across all clusters, coinciding with peak traffic hours in East Asia, Europe, and the east coast of North America. These tasks ran for a total of 4 weeks starting 18 October, 2004. Thus, in this experiment, each path was probed 3,780 times, and the total number of probes was about 652 million. A small number of nodes in the core set became unavailable for extended periods of time due to maintenance or infrastructure changes. A filtering step was applied to the data to purge all the data for these nodes.

A modified all-pairs shortest path algorithm was executed on the data set to determine the shortest paths with one, two, and three intermediate nodes from the core set. For each 4-tuple in the archive, this produced a new 7-tuple `<timestamp, source-id, destination-id, direct RTT, one-hop shortest RTT, two-hop shortest RTT, three-hop shortest RTT>`. The new archive was then split into broad categories based on source and destination continents.

3.3 Evaluation categories

With several gigabytes of data available, it is possible to mine it in a variety of ways. We limit our presentation in this paper to broad categories, and present specific detailed instances when needed for the explanation of unusual observations. The broad categories are based on the continents of the source and destination nodes, motivated by the fact that enterprise websites tend to specify their audience of interest in terms of their continent. The categories are denoted by obvious mnemonics such as AF-NA (indicated in Table 1), denoting that the edge nodes (end users) are in Africa and core nodes (servers) are in North America. In the rest of the paper, we order all tables alphabetically by the source and destination pair for ease of readability.

3.4 Caveats

There are some important consequences of our data collection mechanisms, and the conclusions we draw from them.

– As a consequence of the large measurement testbed, paths are probed once every 2 minutes, to ensure that the maximum rate of ICMP echo requests received by a core node is less than 10 per second, and does not trigger any security related alerts at these data centers. We require three consecutive

probe losses to signal path unavailability, and this implies that one cannot draw conclusions about Internet path failures of duration shorter than 6 minutes.
- Some routers treat small ICMP packets different ly from usual data packets. Thus, normal data transfer along some paths may experience queuing delays not seen by the ICMP probes. Conversely, ICMP packets may be dropped making the destination look unreachable, while the normal data transfer continues unhindered. Our experience indicates, however, that the extent to which ICMP measurements are not representative of end-user performance is small, causing no significant distortion of the results.
- We do not traffic weight our paths due to the lack of a reliable data set on traffic between various Internet nodes; instead, all paths are weighted equally.

4 Performance gains of overlays

This section presents three classes of statistics to provide an understanding of connectivity between different parts of the world. While this section and Section 5 are primarily data driven, their observations will provide intuition for the behaviors identified in Section 6. As mentioned earlier, we evaluate round trip latency as our metric of performance for the applications of interest.

4.1 Performance gains for all paths

We compare the direct and the fastest indirect path for each source destination pair. We divide the data set into buckets based on the continent pair, and the percentage of improvement in the latency of the fastest indirect path (which might be slower than the direct path) as compared to the direct path.

Table 2 shows the percentage of samples that fell in each of the buckets. The rows of the table sum to 100%. As an explanatory example for Table 2, consider the AS-AS row. The bucket $< -10\%$ shows the cases where the best indirect paths are at least 10% slower than the direct path. 15.5% of the AS-AS paths fell in this bucket. The bucket $\pm 10\%$ represents the cases where the best indirect path and the direct path are comparable, in the sense that their latencies are within 10% of each other. 24.7% of the paths in the AS-AS category fell in this bucket. Out of the remaining direct paths, 23.4% saw a marginal (10-30%) improvement, 13.2% of the paths saw significant (30-50%) improvements, and 23.2% of the paths saw large latency reductions of a factor of two or better from the alternate paths found by the overlay.

Prior work in routing overlays has concentrated on intra North America paths, corresponding to the NA-NA row in Table 2. These prior results underestimate the gains in some parts of the world, and overestimate them in some others. For example, the categories of AS-AS, EU-EU and OC-AS have higher percentages in the $> 50\%$ column, while all other inter-continent pairs have lower percentages. For the latter, the common case is that for most direct paths, the

best indirect path is either comparable in RTT, or only marginally faster - the sum of the percentages in the columns $\pm 10\%$ and $10 - 30\%$ ranges from 79% to 96%. This is especially true for inter-continental paths not involving Asian endpoints. The high values of the first and the last columns for the categories

Category	$< -10\%$ (Slower)	$\pm 10\%$ (Comparable)	$10 - 30\%$ (Marginal)	$30 - 50\%$ (Significant)	$> 50\%$ (Large)
AF-AS	4.0	44.5	44.2	5.7	1.6
AF-EU	0.6	69.3	18.1	9.7	2.3
AF-NA	0.0	74.2	21.6	3.5	0.6
AS-AS	15.5	24.7	23.4	13.2	23.2
AS-EU	0.9	33.9	45.5	12.5	7.2
AS-NA	0.1	43.2	42.4	7.6	6.7
CA-AS	0.0	40.5	53.5	4.6	1.4
CA-EU	1.4	53.2	42.3	2.5	0.7
CA-NA	1.7	44.1	41.3	11.2	1.8
EU-AS	0.6	24.5	63.8	7.8	3.2
EU-EU	10.5	36.4	30.5	12.6	10.0
EU-NA	0.0	50.6	45.1	3.3	0.9
NA-AS	0.0	34.0	57.9	5.4	2.6
NA-EU	0.1	43.1	51.1	4.4	1.4
NA-NA	2.4	34.7	39.0	15.0	9.0
OC-AS	6.1	38.9	18.9	22.9	13.2
OC-EU	0.0	60.4	35.1	3.9	0.7
OC-NA	0.0	66.7	25.6	6.3	1.4
SA-AS	0.1	43.1	47.9	5.5	3.4
SA-EU	0.4	66.1	28.9	2.3	2.2
SA-NA	0.9	55.1	35.1	5.7	3.3

Table 2. Histogram of RTT reduction percentages

AS-AS and EU-EU indicates the presence of many cases of pathological routing between ISPs in these continents. A nontrivial number of AS-AS paths is routed through peering locations in California. A specific example can be provided from our snapshot of traceroutes. Consider the path between Gigamedia, Taipei and China Telecom, Shanghai. All the paths in our snapshot that originated at Gigamedia, Taipei and ended at other locations in Asia went via California, except the path to China Telecom, Shanghai, which went directly from Taipei to Shanghai. The Taipei-Shanghai path thus falls in the first column, since all the alternates are very convoluted. At the same time, all the paths that originate in Gigamedia, Taipei and end in other locations in Asia fall in the last column, since their direct routes are very convoluted, but there exists a path via China Telecom, Shanghai, which is more than 50% faster. This example provides the intuition that categories that have a high value in the first column also tend to have a high value in the last column. The same case holds for the OC-AS category. For

the EU-EU category, a similar pattern is seen due to a small number of direct paths going via New York.

4.2 Performance gains for poor paths

The preceding section analyzed the potential benefits of overlays across all paths for a continent pair. However, the benefits provided by overlays in minimizing the worst case path latencies are especially interesting. In this second set of statistics,

Category	50th percentile			90th percentile		
	Direct (ms)	Fastest (ms)	Benefit(%)	Direct (ms)	Fastest (ms)	Benefit (%)
AF-AS	350	290	17	740	700	5
AF-EU	150	120	20	620	620	0
AF-NA	200	180	10	560	550	2
AS-AS	230	110	52	590	350	41
AS-EU	320	260	19	500	360	28
AS-NA	230	200	13	470	280	40
CA-AS	230	200	13	300	250	17
CA-EU	160	140	12	200	170	15
CA-NA	90	70	22	130	100	23
EU-AS	300	260	13	390	300	23
EU-EU	30	30	0	80	60	25
EU-NA	130	120	8	190	160	16
NA-AS	190	160	16	260	210	19
NA-EU	130	110	15	180	150	17
NA-NA	50	40	20	90	70	22
OC-AS	200	140	30	340	220	35
OC-EU	330	300	9	400	330	17
OC-NA	220	200	9	280	230	18
SA-AS	320	280	12	470	340	28
SA-EU	230	210	9	290	250	14
SA-NA	160	150	6	240	190	21

Table 3. RTT reduction percentages for median and poor paths

we compare the RTT reduction benefit as enjoyed by the paths typical of a given continent pair category with that seen by the paths that have high RTT for that category. We divided the data set for each category into 10 millisecond buckets based on the RTT of the direct path, and determined the improvements provided by the fastest path over the direct path for the 50th and 90th percentile of all paths between a continent pair. Table 3 shows the comparison of the benefits seen by the typical (median) and the poor paths in each category. For the typical paths, the RTT reduction benefit exceeds 20% only for AS-AS, OC-AS and CA-NA out of the 21 categories. Comparatively, the poor paths see a benefit over 20% for half of the categories. The important categories of AS-AS, AS-NA, and

EU-EU show significant improvements for the poor paths, while, in contrast for paths originating from Africa the RTT at the 90th percentile is both high and not helped with the overlay. For the AS-AS category, both the typical and poor paths see significant improvement via the overlay, where the improvement is even greater for the typical paths. For all of the other categories, the poor paths benefit more from the overlay, compared to the typical paths.

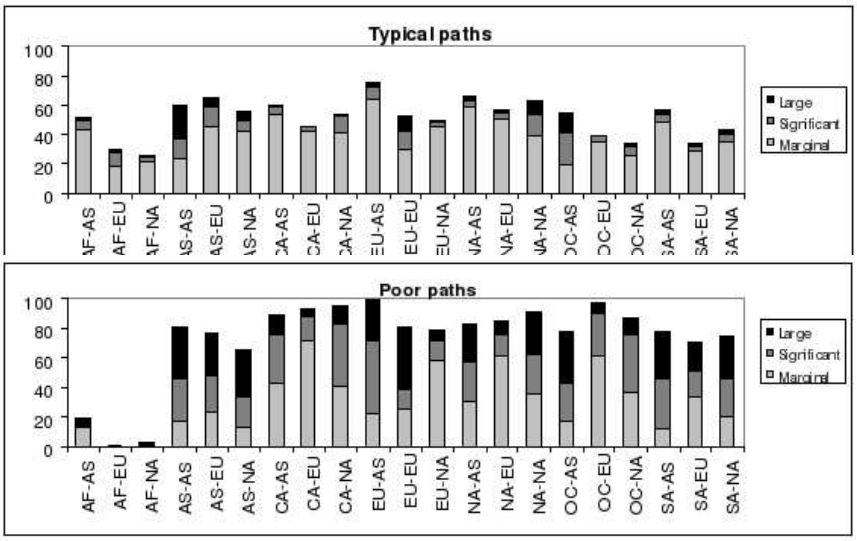

Fig. 1. RTT reduction for all and for poor paths

In our third set of statistics, we focus on all the poor paths whose direct latency ever exceeded the 90th percentile direct latency shown in Table 3. (Note that the data in Table 3 for the 90th percentile section does not include paths in higher buckets, but the following analysis does). We then recomputed the histogram of RTT reduction (such as the one in Table 2) only for those paths. The distribution of the samples in this table is dramatically different compared to Table 2. The last three columns (at least 10% improvement) carry the most weight of each category. For easier comparison with Table 2, this data is presented graphically in Figure 1. The first bar-chart in Figure 1 shows the last three columns of Table 2 (all paths). The second bar-chart shows the same for the poor paths. It is seen that when focusing on poor paths, over 80% of the categories see at least marginal benefits, while 67% of the samples see significant or large benefits. Some categories do deviate from this observation in the figure. For example, even poor paths originating in Africa do not derive much help from an overlay.

5 Availability gains of overlays

Improving the availability of the transport layer by routing around route failures is the second important functionality of overlays that we consider in this paper.

In Section 5.1, we continue on the theme of the previous section, and present statistics about how often direct-path failures were encountered during our measurements, and what fraction of times at least one alternate path was functional during this failure. This allows us to form an expectation about availability gains that an overlay can provide. Section 5.2 focuses on the particularly poor paths in terms of availability, and examines the benefit of using overlays for these paths.

5.1 Availability gains for all paths

Our ICMP echo request tasks report a special latency value when the sending node gets no reply from the destination. A single lost packet does not necessarily indicate path unavailability, but rather a congested or lossy path. However, since congestion is a transient event lasting only in the range of seconds or less, the loss of a sequence of consecutive packets is considered a strong indicator of path unavailability. As mentioned in Section 3.2, a ping was sent every 2 minutes between each edge-node, core-node pair, (which given the number of nodes, keeps the rate that any node responds to pings to within a nominal value of 10 pings/sec). We consider a path to be unavailable if three or more consecutive pings are lost. Akella et al. in [2] use the same definition, where the pings were sent at 1 minute intervals. The alternative scenario that three consecutive pings are each lost due to random congestion occurs with a probability of order 10-6, assuming independent losses in two minute epochs with a probability of order 1%. We consider the unavailability period to start when the first lost ping was sent and to end when the last of the consecutively lost pings was sent. This is likely a conservative estimate of the length of the period. We filtered out all measurements originating from edge nodes in China for our availability analysis. Their failure characteristics are remarkably different from all other Internet paths as a consequence of firewall policies applied by the Chinese government.

A simple analysis of path failures was carried out to compute path availabilities with and without the alternate paths. Figure 2 shows the results of this analysis. Note that this figure plots the percentage unavailability for different continent pairs, in order to clarify the impact of overlays on improving reliability. The unavailability of the direct paths ranges from 0.03% to 0.83%. Asia has the poorest availability: nine of the ten lowest availability categories have an endpoint in Asia. In the presence of an overlay, the availability of the transport goes up by 0.3-0.5% for most categories. The low-availability categories involving Asia show dramatic availability improvements.

5.2 Availability gains for poor paths

As with section 4.2, we now analyze the benefits provided by overlays for paths with abnormally low availability, as an attempt to understand the benefits of

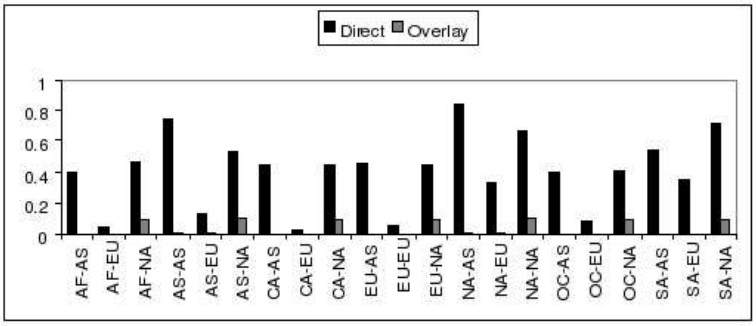

Fig. 2. Reduction in path failure percentages with the overlay

Category	% paths with 30% failures	Failure % no overlay	Failure % overlay
AF-AS	4.5	25.8	0
AF-EU	1.7	8.8	0
AF-NA	0.6	36.2	0
AS-AS	2.7	31.4	0
AS-EU	1.5	9.8	0
AS-NA	0.4	30	0
CA-AS	3.5	28.2	0
CA-EU	1.6	10.9	0
CA-NA	0.5	30.3	0
EU-AS	3	30.1	0
EU-EU	0.9	10.8	0
EU-NA	0.4	30.1	0
NA-AS	2.7	32.3	0
NA-EU	0.4	13.2	0
NA-NA	0.2	40.2	0
OC-AS	3.1	30.8	0
OC-EU	1.4	10.7	0
OC-NA	0.4	29.3	0
SA-AS	3.3	28.8	0
SA-EU	2.2	9.5	0
SA-NA	0.8	23	0

Table 4. Availability statistics for poor paths

overlays in alleviating the worst case. It is commonly understood that a small number of paths contribute to a large number of path failures on the Internet. As evaluated in [13], 3% of Internet paths give rise to 30% of failures. We identified a similar pattern in our data as shown in Table 4. We see that about 3% of the paths caused 30% of the failures, and that 10% of the paths gave rise to 50% of the failures.

We identified the paths in each category that gave rise to 30% of the failures, and re-ran the availability analysis of Section 5.1 for only these paths, to examine the benefits of overlays for the most unavailable paths in each category. This data is also shown in Table 4.

A failure rate higher than 20% on a path is indicative of some specific chronic trouble with the path, rather than random, transient failures or short-lived congestion. Almost all these path availability issues disappear with an overlay. This indicates that the overlay was able to route the traffic around the problem. The improvement made to chronically failing paths is the key to availability improvements by the overlay.

6 Achieving the benefits in a practical design

All the analysis presented thus far is for an ideal case, where the network latency measurements at a particular time are used in the calculations for alternate paths, whose performance benefits are then determined using these self same measurements. This analysis can therefore be considered as providing an upper limit on the performance and availability gains that can be expected from an overlay.

Conceptually, one could obtain these benefits by using a brute force approach of flooding the data along all direct and indirect paths. This approach is clearly not practical, and real overlay systems have the following fundamental design constraints:

- Cost and network friendliness constraints prevent the use of all possible indirect paths. Instead, practical algorithms will need to use a policy that sends packets along one path or a small set of judiciously chosen paths.
- The routing decisions at any given instant will need to be based on measurements of events in the past, and used for a window of time in the future.

To meet a performance target, a system designer typically considers (a) the frequency of active probing and the window of persistence of a decision and (b) the number of paths on which copies are sent. A judicious choice balances the overhead and performance gains of each.. This section addresses these constraints in trying to answer how hard it would be to build an overlay that approaches the results in the ideal case.

6.1 Random path selection policy

We first analyze a simple multi-path memoryless random policy to select the subsets of paths used to transmit data, based purely on static information and

not dynamic network data. For each time sample, and source-destination pair, our algorithm picks a set of alternate paths chosen at random from the list of candidates, and compares how effective this overlay is when compared to the ideal. It is natural to expect that this overlay will likely be inferior to the ideal, but our goal is to develop a straw man to validate the importance of intelligence and adaptiveness in overlay selection. We limit this analysis to only three categories of paths: AS-AS, NA-NA, EU-EU.

We did coarse pruning of the initial candidate pool based on geography to avoid picking alternate paths which are guaranteed to be highly detrimental to performance. Specifically, for the AS-AS paths, we choose intermediate regions only in AS and the west coast of North America. For EU-EU paths we implement this choice a little more coarsely by choosing only intermediate nodes in EU. We implement the same pruning algorithm in NA with a slight modification: if both nodes are on the east coast or both on the west coast of North America, we pick candidates from nodes in these respective locations.

We then re-compute the results in Table 2 and Figure 2 with random-path subsets of sizes 1, 2, 3 and 5. An interesting result is that a random policy with a set size of 3 provides availability gains comparable to the ideal case for paths in the categories NA-NA and EU-EU. AS-AS paths, however, require a set size of 5 to achieve similar gains. Random path selection was unable to provide any significant performance gains, though. In fact, over 90% of the samples provided no benefit - they were either over 10% worse than the direct path, or within 10% of the direct path (the first two columns of table 2).

The success of a random selection algorithm in providing near optimal availability is significant. This result substantiates the fact that the Internet offers very good path diversity, and generally has low rates of failure. The higher subset size required for paths with both end points in Asia is most likely a consequence of a higher rate of path failures in that region.

The failure of this policy in improving performance, however, suggests that careful path selection is very important in building overlays for performance gains.

6.2 Deterministic path selection policy

In the rest of this section, we analyze deterministic selection policies with a goal of understanding reasonable parameter choices for a performance optimizing overlay. A natural notion would be to examine algorithms that make decisions based on measurements at time t and use it at a future time t'. We parameterize these overlays with two variables, the number of paths on which to send data, and the time into the future for which we hold decisions fixed.

Formally, we denote an overlay that chooses the κ best performing paths (one of which may be direct) and predicts for up to τ minutes into the future as $O(\kappa, \tau)$. An overlay with κ paths would then send traffic along the chosen κ paths, and therefore achieve the performance of the best of those κ paths at the cost of a κ-fold amplification in traffic.

Stability of optimal paths To the extent that an overlay for performance selects a subset of paths to use, it will deviate from optimality as a result of variations in path latencies that cause a reordering of the best paths. Paths tend to fall into two categories:

1. The best paths from an edge region to an origin are quite persistent, and do not change, regardless of variations in the round trip times of all paths.
2. RTT variations of the paths over time cause a significant reordering of the best paths, and changes in the optimal paths.

Paths in the first category do not require a very dynamic overlay design for selection of alternate paths for performance improvement. For examples of paths in the first category, consider the path from Pacific Internet, Singapore to AboveNet, London. The direct path, which hops from Singapore through Tokyo, San Francisco, Dallas, and Washington D.C. to London takes approximately 340 msec. However, there exists an alternate path through an intermediate node in the ISP Energis Communications in London. The path between Pacific Internet, Singapore and Energis, London is one hop long (possibly a satellite link), and has a latency of 196 ms. The subsequent traversal from Energis, London to AboveNet, London takes just 2 ms. The alternate path is therefore faster than the direct path by over 140 ms, or 41.2%. While the difference in latencies are quite glaring in the above pathological example, it is indeed common to find several cases where the differences between the direct and indirect paths are significant enough that transient latency variations do not affect this ordering. In these cases, as one can imagine, a performance optimizing overlay need not be very dynamic in its path selection algorithm to achieve the ideal case performance gains.

We formally examine the extent of the variation across paths by computing a statistic called churn, to measure how much the sets of optimal paths at two different time instants vary. Formally, for a given pair of nodes, $Churn(\kappa, \tau)$ is defined as the average over time of the value $|S(\kappa, \tau) - S(\kappa, t + \tau)|/\kappa$, where $S(\kappa, \tau)$ is the set of the κ best performing paths between those nodes at time t. $Churn$ is a number between 0 and 1, will be 0 for paths with a persistent set of best paths, and tend to be closer to 1 for paths with a fast changing set of best paths. Intuitively, for a given set size, $Churn$ will increase as we use sets based on estimates further in the past; and for a given time window, $Churn$ will decrease as we increase the set size.

For each path, we compute the value of $Churn(t, t + \tau)$ for all valid values of t and different values of τ, and determine the average value for a given window size τ. We then bucket these paths into the same continent based categories as before, and compute these values for all overlays $O(\kappa, \tau)$ where κ takes all values between 0 and 5, and τ is 2, 4, 6, 8, or 10 minutes.

We found that the majority of paths have values of $Churn$ larger than 10%, even when selecting up to the top 5 best performing paths and using this prediction only 2 minutes into the future. At first blush, this seems to cast a bleak light on the prospects of constructing a performance optimizing overlay that provides significant gains.

However, *Churn* does not provide any measure of how much worse the elements leaving the old set were compared with the elements in the current set. We obtain a measure of the extent of the performance degradation of changing sets by computing a relaxed value called *RelaxChurn* where we count only exiting elements whose current performance is more than 10% worse than the worst performing element in the current set, i.e. their presence in the current set would not worsen the performance by more than 10% (and, in fact, might not impact the performance at all depending on the other elements in the set).

Again, *RelaxChurn* is a number between 0 and 1, with values closer to 0 indicating a higher degree of constancy in the ordering of paths (after applying relaxation). We also expect *RelaxChurn* for a path to be less than *Churn* for a given set size and time window. Very hearteningly, *RelaxChurn* values are less than 10% on average for over 80% of paths in most categories. This indicates that a path selection algorithm that makes predictions into the future based on current measurements, can achieve performance close to the ideal.

We use the figure of 10% as a threshold to classify paths as low-churn or high-churn. Figure 3 shows the percentage of paths that have Churn and RelaxChurn of less than 10% for $\kappa = 1$ and $\tau = 2$ minutes. We have excluded the churn numbers for higher values of κ and τ here to limit the amount of data presented here. As an explanatory example, consider the bar for the category NA-NA. It shows that 27.3% of the paths in this category are low *Churn* paths, and 74.3% of the paths are low *RelaxChurn* paths.

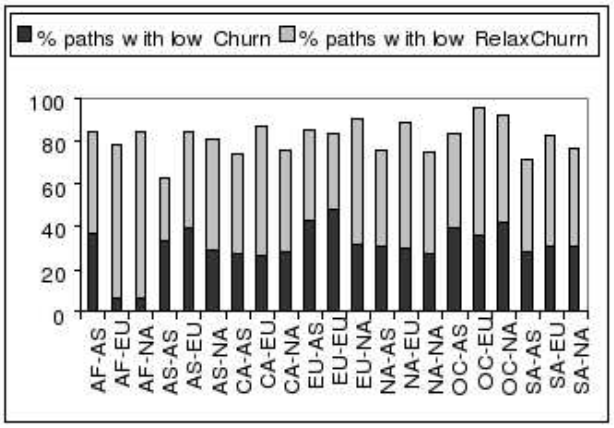

Fig. 3. Percentage of paths with low *Churn* and *RelaxChurn* for $\tau = 2$ minutes and $\kappa = 1$

Note that paths with both the end points in Asia tend to have a marginally higher value of *RelaxChurn* (only 63% AS-AS paths are low-churn paths) compared to all other paths. It thus seems like the potential higher performance benefits obtainable for AS-AS paths come at a higher cost in terms of network measurement.

Performance gains of a predictive overlay The previous analysis examined stability using purely structural properties. In this section, we examine how far sets chosen in the past deviate from the ideal sets at any time instant. More precisely, for a given source destination pair, we compute the performance loss suffered by the choice of an overlay $O(\kappa, \tau)$ over the ideal decision at the given time. We expect this number to decrease with increased set size for a given window, and increase with an increased window size for a given set size. Note that this measure holds overlays to a higher standard, as the ideal path at a given time is at least as fast as the direct path.

A natural case to examine in some detail would be $\kappa = 1$. This corresponds to just using the best path choice in future iterations. Table 5 in the second and third columns shows this data for $\tau = 2$ and 10 minutes. Again, we limit the amount of data presented here for ease of exposition, by showing only these two values for τ. As an explanatory example, consider the NA-NA category. It shows that when using $\tau = 2$ minutes, 71.6% of the paths came within 10% of the optimal latency for that observation. Even when using stale data, with $\tau = 10$ minutes, 69.6% of the paths managed to achieve the same result.

Paths originating in Asia again show a greater deviation from optimality than paths originating in Europe, whereas paths originating in North America span the full range of deviations. Given that the performance gains of the family of overlays $O(1, \tau)$ do not seem adequate everywhere, we then explored overlays of increasing sizes. Table 5 in the second, fourth and fifth columns shows the percentage of paths that come within 10% of the optimal latency. As an explanatory example, consider the category NA-EU. It shows that 82.3% of the paths came within 10% of the optimal when choosing $\kappa = 1$. Increasing κ to 2 enabled approximately 13.1% more paths to achieve the same result. Increasing κ to 3 has only a marginal benefit to the remaining paths, and only 1.8% more paths achieved the result with this value of κ. From Table 5, we immediately see that choosing $\kappa = 2$ provides disproportionately high gains over choosing $\kappa = 1$, and the marginal benefit of choosing $\kappa = 3$ is much lower.

In fact, other than paths with their destination in Asia, over 90% of all paths are within 10% of the ideal performance when selecting $\kappa = 2$, and this fact remains true even with increasing τ.

The results in Table 5 suggest the potential for an adaptive multi-path scheme where, for a given source destination pair and given time, either 1 or 2 paths are used. For example, 95.4% of all NA-EU paths are within 10% of optimal for overlays with $\kappa = 2$. Combining this with the fact that 82.3% of these paths require only one choice to come within the same limits, it is conceivable that an adaptive multi-path strategy could use two paths only for the excess 13.1% of paths, for an average overhead of just 1.09 paths.

Paths with both end points in Asia and Europe are not lagging too far behind, however, and around 85% of paths get a corresponding benefit. For example, the proportion of AS-AS paths within 10% of optimal jumps from 62.44% to 84.57% when going from $\kappa = 1$ to $\kappa = 2$ (for a weighted average set size of 1.31). Getting to a 90% number for the other paths, however, requires $\kappa = 3$. Although Table 5

Category	Percentage of paths			
	$\kappa = 1$ $\tau = 2$	$\kappa = 1$ $\tau = 10$	$\kappa = 2$ $\tau = 2$	$\kappa = 3$ $\tau = 2$
AF-AS	74.0	72.6	88.8	94.6
AF-EU	83.0	82.0	95.4	97.4
AF-NA	81.5	79.8	95.4	97.1
AS-AS	62.4	59.5	84.6	89.4
AS-EU	76.2	74.1	92.2	94.5
AS-NA	74.8	71.6	94.0	96.0
CA-AS	57.8	56.0	90.0	94.9
CA-EU	77.4	76.4	97.6	98.6
CA-NA	68.6	66.6	96.3	98.2
EU-AS	74.4	72.3	88.4	92.8
EU-EU	80.1	78.1	91.6	93.1
EU-NA	83.0	82.2	94.7	96.2
NA-AS	68.1	66.2	88.8	93.7
NA-EU	82.3	81.3	95.4	97.2
NA-NA	71.6	69.6	92.0	95.0
OC-AS	75.3	74.3	89.2	92.4
OC-EU	81.0	79.6	97.0	98.4
OC-NA	77.9	76.4	97.1	98.2
SA-AS	52.9	50.2	84.4	90.6
SA-EU	77.6	76.3	91.3	92.9
SA-NA	70.5	68.4	90.6	92.7

Table 5. Percentage of paths within 10% of the optimal latency

Category	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	Ideal
AF-AS	97.47	99.34	99.93	100.00
AF-EU	99.17	99.84	99.98	100.00
AF-NA	99.04	99.73	99.87	99.99
AS-AS	96.43	99.04	99.89	100.00
AS-EU	99.04	99.73	99.94	100.00
AS-NA	98.98	99.78	99.92	99.98
CA-AS	95.56	99.23	99.86	100.00
CA-EU	98.69	99.89	99.98	100.00
CA-NA	98.89	99.79	99.99	100.00
EU-AS	97.13	99.34	99.88	100.00
EU-EU	98.99	99.80	99.95	100.00
EU-NA	98.83	99.66	99.87	100.00
NA-AS	96.79	99.33	99.78	100.00
NA-EU	99.13	99.85	99.98	100.00
NA-NA	99.04	99.73	99.87	100.00
OC-AS	96.43	99.07	99.89	100.00
OC-EU	98.92	99.85	99.96	100.00
OC-NA	98.78	99.77	99.84	100.00
SA-AS	94.76	99.01	99.66	100.00
SA-EU	98.24	99.74	99.83	100.00
SA-NA	97.99	99.69	99.83	100.00

Table 6. Availability variation with κ, for $\tau = 2$ minutes

shows results for $\tau = 2$ minutes for $\kappa = 2$, these values remain relatively stable for higher values of τ between 2 and 10 minutes (similar to the case of $\kappa = 1$). This implies that increasing the rate of probing does not lead to gains in latency for a significantly higher number of paths. We expand on these results in Section 6.3

Interestingly, overlays designed for high performance show reduced availability as compared to the ideal situation. Table 6 shows the extent to which the best predictive overlay for $\tau = 2$ minutes deviates from the ideal for several values of κ. As the table shows, increasing κ gets the overlay closer to the ideal, but these overlays are still inferior to the random (optimal) overlays described in Section 6.1. This is because, as illustrated in earlier examples in this paper, better performing paths are typically constrained to share a small set of common links leading to lower path diversity. In contrast, most links on the Internet are available most of the time, and hence random selection of paths provides sufficient diversity to guarantee high availability. The optimality of random paths suggests that a judicious hybrid of the two strategies can construct overlays that approach the ideal, both for performance and availability.

6.3 Long term persistence

The data in Section 6.2 indicates that the benefits of overlays are only mildly sensitive to the value of τ, at least in the range of 2 to 10 minutes. In this section, we explore the time sensitivity of predictive overlays by using some extreme cases. Our daily 1.5 hour samples are separated by between 4 and 11 hours. We used overlays based on measurements in one 1.5 hour sample, and evaluated their performance on the next sample. While it is entirely possible that the overlay might have been suboptimal in the intervening time period, a result showing that a large number of paths is either very close or very far away from optimal would be indicative of the long term dynamics of overlays. In fact, we see that around 87% of NA-NA, and 74% of AS-AS paths are within 10% of ideal even with these long term predictions. These statistics point to a high degree of consistency in the relative performance of alternative paths between a source-destination pair, for most pairs.

6.4 Optimizing the outliers

Table 5 shows that most categories have about 90% of paths achieving within 10% of the ideal by using overlays with $\kappa = 2$ (for $\tau = 2$ minutes), and that overlays with $\kappa = 3$ increase this number by between 2 and 7 percentage points. One might naturally assume that increasing κ would gradually increase the number of paths that are close to ideal. However, as Figure 4 shows, this is not the case, and the proportion of paths within 10% of ideal performance quickly reaches saturation for values of κ around 5 or 6. This is an indication that there is a small number of paths with high short term performance variations, and it is difficult for a predictive overlay to optimize these paths. This is an interesting counterpoint to the results of Section 6.3 which establish a high degree of persistence for the majority of paths.

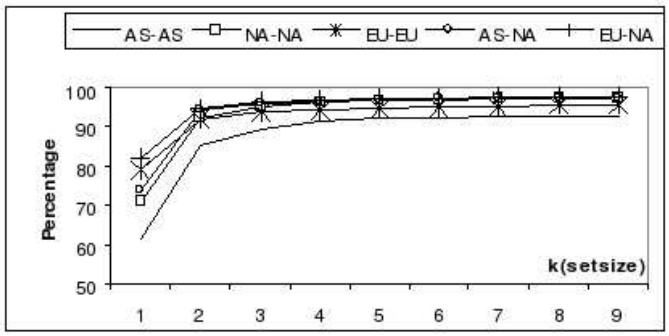

Fig. 4. CDF of latency benefits with κ

7 Concluding remarks

In this paper, we quantified the performance and availability benefits that can be extracted from the current Internet by routing overlays, and how it differs from continent to continent or for transcontinental paths. Our analysis is of a global nature, with the evaluation network containing over a thousand nodes in 6 continents, 77 countries, and 630 cities. We see that the performance and availability benefits that can be provided by overlays in Asia, and often in Europe, far exceed the hitherto observed values for paths in North America. We also analyze practical overlay designs, and show that it is possible to design real overlays that achieve near-ideal performance and availability for a large number of paths.

One question that arises from our work is whether the significant differences in behavior internationally will disappear over time as connectivity and economies improve, or are artifacts of deeper structural issues. Industry experts have reason to believe that the levels of interconnectivity in many international markets are consequences of the economics and traffic patterns, and hence are likely to persist. To quote Farooq Hussain [8], a connections manager experienced in global network relationships, Despite these European changes, patterns in the rest of the world have not changed much over time. Despite initiatives to establish regional network infrastructure in the Americas for example, the main flow of traffic from countries there is to the US and inter-regional traffic volumes don't justify the building of networks to serve regional needs. Similarly, in Asia, efforts to send European bound traffic directly there instead of via the US, though established, are not competitive with trans-pacific routes. Also, many countries have found that the inter-regional traffic within Asia doesn't justify connections between countries in the region It's hard to see how the Americas or Africa can deploy routes with the equivalent competition since they also do not operate as an economic group in the manner of the European community".

7.1 Future Work

Our present study suggests several directions that warrant further research about routing overlays.

First, as mentioned in the introduction, the present study of overlay routing from edge nodes to core nodes is tailored for traffic between clients and enterprise servers. A natural analogous study would be tailored for peer-to-peer traffic and consider edge node to edge node routing.

Second, as discussed in section 6.2, there is significant potential to reduce the overhead associated with a multi-path routing overlay by adaptively adjusting the number of redundant paths, where typically just one path is sufficient, while two, and less frequently, three paths are used as needed. The design of the adaptive scheme is an interesting direction for future work.

Third some research to characterize and understand the cause of the long-term persistence of the good performance obtained from given overlay paths, presented in Section 6.3,is also warranted. Part of the cause is likely sparse inter-regional infrastructure as cited above in the conclusions [8], and part of it is likely due to some ISP's policies. An ISP may have poor peering with others, or it may choose to engineer traffic in a way that compromises performance for given end-users. Overlay routing can in effect override an ISPs desired policy. Research in this area has examined the impact on network-wide measures such as average latency when individual end-systems can select routes based individual criteria, termed selfish routing [17], as well as simulation studies to examine the impact of traffic oscillations [12]. One can envision a tussle [7] between ISPs and overlay-routing service providers, though the volume of traffic being re-routed via overlay networks would need to grow beyond the noise in an ISPs demand forecast, before the engineering conflict is readily apparent. What will likely be apparent sooner will be the revenue obtained by commercial overlay-routing service providers. This will give an incentive for sets of ISPs to collaborate and offer collectively an enhanced routing service to customers who are willing to pay for it. The future industry structure of commercial overlay routing is unclear and would be interesting to understand as part of a research agenda.

Bibliography

[1] Akamai Technologies, Inc. http://www.akamai.com.

[2] A. Akella, J. Pang, B. Maggs, S. Seshan, and A. Shaikh. A comparison of overlay routing and multihoming route control. In *Proc. ACM SIGCOMM*, pages 93–106, Portland, OR, Aug. 2004.

[3] D. G. Andersen. *Improving End-to-End Availability Using Overlay Networks*. PhD thesis, MIT, 2005.

[4] D. G. Andersen, H. Balakrishnan, F. Kaashoek, and R. Morris. Resilient Overlay Networks. In *18th ACM SOSP*, Banff, Canada, October 2001.

[5] Ibm to axe 13,000 jobs worldwide. http://news.bbc.co.uk/1/hi/business/4515101.stm.

[6] CAIDA. http://www.caida.org.

[7] D. Clark, J. Wroclawski, K. Sollins, and B. Braden. Tussle in cyberspace: defining tomorrow's internet. In *ACM SIGCOMM*, 2002.

[8] The Cook Report on the Internet, Vol XI, Nos 5-7 Aug-Oct 2002. http://cookreport.com/11.05-6.shtml.

[9] Detour. http://www.cs.washington.edu/research/networking/detour/.

[10] FCW Media Group – Battlefield Communications. http://www.fcw.com/article88262.

[11] K. P. Gummadi, H. Madhyastha, S. Gribble, H. Levy, and D. Wetherall. Improving the reliability of internet paths with one-hop source routing. In *OSDI*, San Diego, CA, 2003.

[12] R. Keralapura, N. Taft, C. Chuah, and G. Iannaccone. Can isps take the heat from overlay networks? In *ACM SIGCOMM Workshop on Hot Topics in Networks (HotNets)*, 2004.

[13] A. Markopoulou, G. Iannaccone, S. Bhattacharyya, C.-N. Chuah, and C. Diot. Characterization of failures in an ip backbone. In *IEEE Infocom*, Hong Kong, 2004.

[14] NIMI. http://ncne.nlanr.net/nimi/.

[15] V. Paxson, J. Mahdavi, A. Adams, and M. Mathis. An Architecture for Large-Scale Internet Measurement. *IEEE Communications*, August 1998.

[16] Planetlab. http://www.planet-lab.org/.

[17] L. Qiu, Y. R. Yang, Y. Zhang, and S. Shenker. On selfish routing in internet-like environments. In *ACM SIGCOMM*, 2003.

[18] RON. http://nms.csail.mit.edu/ron/.

[19] S. Savage, A. Collins, E. Hoffman, J. Snell, and T. Anderson. The end-to-end effects of internet path selection. In *Proc. ACM SIGCOMM*, pages 289–299, Cambridge, MA, 1999.

[20] Zona Research. The need for speed II. Zona Market Bulletin 5 (Apr. 2001).