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

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

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 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. Regarding practical considerations in constructing overlay routes, we show that an 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, and show that this is achievable with surprisingly persistent routing choices.

1 Introduction

There has been much recent work [3, 9, 18] on improving the performance and availability of the Internet using routing overlays. Business trends such as outsourcing and workforce consolidation [5], as well as stringent requirements for applications such as government communications [10], necessitate that these performance and availability improvements are obtained not just within a single country or small group of countries, but globally. While existing 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.

*Supported in part by an NSF award under grant number CNS-0519894.

– Overlays used in these studies have a footprint primarily in North America. However, it is well known [8] that network interconnectivity and relationships in Europe and Asia are different than the continental United States.
– 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.

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 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.

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.

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 provides useful cross validation for the currently deployed testbeds such as PlanetLab [15] and RON [18], and indicates that, while these deployments provide qualitatively similar data for the commercial Internet in North America, they do not capture the global diversity of network topology, especially in Asia.
– We show that randomly picking a small number of redundant paths (3 for Europe and North America, and 5 for Asia) achieves availability gains that approach the optimal. Additionally, we demonstrate that for reasonable probing intervals (say, 10 minutes) and redundancy (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.

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 describes our testbed and how the measurement data is collected. Section 4 provides detailed metrics on the ideal availability and performance gains that can be achieved by overlays in a global context. Section 5 addresses issues in real overlay design, and explores structural and temporal properties of practical overlays for performance and availability. In the interests of space, we present details only for the performance results, and refer the reader to the technical report [17] for expanded availability results.

2 Related Work

There have been many measurement studies of Internet performance and availability, for example, the work at the Cooperative Association for Internet Data Analysis (CAIDA) [6], and the National Internet Measurement Infrastructure (NIMI) [13, 14]. Examples of research routing overlay networks are the Resilient Overlay Networks project at MIT [18] and the Detour project at U. Washington [9]. A commercial routing overlay is offered by Akamai Technologies [1].

Andersen *et al.* [4] present the implementation and performance analysis of the routing overlay called Resilient Overlay Networks (RON). They found that their overlay improved latency 51% of the time, which is comparable to the 63% we obtain for paths inside North America (details in [17]). Akella *et al.* [2] investigate how well a simpler route-control multi-homing solution compares with an overlay routing solution. Although the focus of that study is different than the present paper, it includes results for a default case of a single-homed site, and the authors find that 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 the 110 node, intra-North-America case herein, where we find that the overall latency improvement is approximately 21%, although the improvement varies significantly across different continent pairs. (details in [17]). Savage *et al.* [19] used data sets of 20 to 40 nodes and found that for roughly 10% of the host pairs, the best alternative has 50% lower latency. We obtain the comparable value of 9% of host pairs for the case of intra-North America nodes, though again significantly disparate results for other continent pairs. 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.

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

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 of the 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 (this keeps the rate of requests at a core node to less than 10 per second). Each task lasted for 1.5 hours. If a packet was lost (no response received within 10 seconds), then a special value was 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. We obtained an archvie of 7-tuples `<timestamp, source-id, destination-id, direct RTT, one-hop shortest RTT, two-hop shortest RTT, three-hop shortest RTT>`. The archive was split into broad categories based on source and destination continents.

We consider a path to be unavailable if three or more consecutive pings are lost. Akella *et al.* [2] use the same definition, where the pings were sent at one 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, and implies that we only draw conclusions about Internet path failures of duration longer than 6 minutes.

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.

3.3 Evaluation

In the interests of space and clarity, we limit our presentation in this paper 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 AS-NA (indicated in Table 1), denoting that the edge nodes (end users) are in Asia and core nodes (servers) are in North America. Table 2 is restricted only to paths with source and destination nodes in AS, NA, and EU, and the reader is referred to [17] for the complete details.

4 Availability and Performance Gains of Overlays

This section presents brief statistics to provide an understanding of connectivity between different parts of the world, and to develop intuition for the behaviors described in Section 5. We present only a summary here in the interests of space, and refer the reader to the technical report [17] for details.

In the presence of an overlay, the availability of the transport goes up by 0.3-0.5% for most categories, though routes in Asia see gains larger than 3.25%. Asia has the poorest availability: nine of the ten lowest availability categories have an endpoint in Asia. Additionally, the key to availability improvements by the overlay is the improvements made to chronically failing paths [17].

About $4 - 35\%$ of all paths see improvements of over 30% in round trip latency (performance). For the subset of poorer performing paths (those whose direct latency exceeds the 90^{th} percentile for given continent pair), about 67% see over 30% improvement. Additionally, high numbers of paths see over 50% improvement for the AS-AS and EU-EU categories, which indicates the presence of many cases of pathological routing between ISPs in these continents. A non-trivial number of AS-AS paths is routed through peering locations in California, for example, the path between Gigamedia, Taipei and China Telecom, Shanghai. All the traceroutes 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 sees little or no improvement with an overlay, since all the alternatives are very convoluted. At the same time, all the paths that originate in Gigamedia, Taipei and end in other locations in Asia see *large* improvements, since their direct routes are very convoluted, but there exists a path via China Telecom, Shanghai, which is more than 50% faster.

5 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 bound on the performance and availability gains that can be expected from an overlay. This section addresses practical constraints in trying to answer how hard it would be to build an overlay that approaches the results in the ideal case.

We first evaluate a simple multi-path memoryless policy to randomly select the subsets of paths used to transmit data, based purely on static information. 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. Surprisingly, random selection is successful in providing near optimal availability, substantiating the fact that the Internet offers very good path diversity, and generally has low rates of failure. The policy, however, fails in improving performance, suggesting that careful path selection is very important in building overlays for performance gains.

5.1 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 a given time, and use them at a future time. We parameterize these overlays with two variables, the number of paths on which to send data (κ), and the time into the future for which we hold decisions fixed (τ).

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:

- The best paths from an edge cluster to an origin are quite persistent, and do not change, regardless of variations in the round trip times of all paths.
- 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%.

We systematically 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, $Churn_t(\kappa, \tau)$ is defined as $|S(\kappa, t) - S(\kappa, t + \tau)|/\kappa$, where $S(\kappa, t)$ is the set of the κ best performing paths between those nodes at time t. $Churn(\kappa, \tau)$ for a node pair is then computed as an average of $Churn_t$ over all valid values of 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. 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. However, a relaxed measure, $RelaxChurn$, that counts only exiting elements whose current latency is higher than 110% of the latency of 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%) produces values 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.

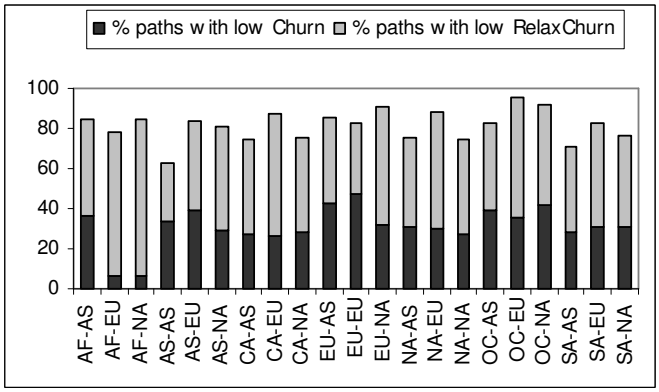

Fig. 1. Percentage of paths with low $Churn$ and $RelaxChurn$ for $\tau = 2$ minutes and $\kappa = 1$

Figure 1 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. 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 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 2 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

Category	Percentage of paths				Category	Percentage of paths			
	$\kappa = 1$ $\tau = 2$	$\kappa = 1$ $\tau = 10$	$\kappa = 2$ $\tau = 2$	$\kappa = 3$ $\tau = 2$		$\kappa = 1$ $\tau = 2$	$\kappa = 1$ $\tau = 10$	$\kappa = 2$ $\tau = 2$	$\kappa = 3$ $\tau = 2$
AS-AS	62.4	59.5	84.6	89.4	EU-NA	83.0	82.2	94.7	96.2
AS-EU	76.2	74.1	92.2	94.5	NA-AS	68.1	66.2	88.8	93.7
AS-NA	74.8	71.6	94.0	96.0	NA-EU	82.3	81.3	95.4	97.2
EU-AS	74.4	72.3	88.4	92.8	NA-NA	71.6	69.6	92.0	95.0
EU-EU	80.1	78.1	91.6	93.1					

Table 2. Percentage of paths within 10% of the optimal latency

of overlays $O(1, \tau)$ do not seem adequate everywhere, we then explored overlays of increasing sizes. Table 2 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 provides only a marginal benefit for the remaining paths, and only 1.8% more paths achieved the result with this value of κ. From Table 2, 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, apart from 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 2 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 2 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 5.1

Interestingly, overlays designed for high performance show reduced availability as compared to the ideal situation. 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.

Persistence The data in Section 5.1 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. In contrast, there is a small number of paths [17] with high short term variations, and it is difficult for a predictive overlay to optimize these paths even with κ going up to 5 or 6.

6 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. The significant differences

in behavior internationally are artifacts of deeper structural issues, and are not expected to disappear over time as connectivity and economies improve. For instance, [8] observes that countries in the Americas, Asia or Africa cannot deploy routes with the equivalent competition, since they also do not operate as an economic group in the manner of the European community.

However, the potential impact of interactions between overlay routing and ISP policies [12, 16] leaves the future industry structures of ISP routing and commercial overlay routing unclear. An initial discussion of these issues can be found in [7].

References

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, B. Lehr, S. Bauer, P. Faratin, R. Sami, and J. Wroclawski. The Growth of Internet Overlay Networks: Implications for Architecture, Industry Structure and Policy. In *33rd Research Conference on Communication, Information and Internet Policy*, Arlington, Virginia, September 2005.
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. NIMI. http://ncne.nlanr.net/nimi/.
14. V. Paxson, J. Mahdavi, A. Adams, and M. Mathis. An Architecture for Large-Scale Internet Measurement. *IEEE Communications*, August 1998.
15. Planetlab. http://www.planet-lab.org/.
16. L. Qiu, Y. R. Yang, Y. Zhang, and S. Shenker. On Selfish Routing in Internet-Like Environments. In *ACM SIGCOMM*, 2003.
17. H. Rahul, M. Kasbekar, R. Sitaraman, and A. Berger. Towards Realizing the Performance and Availability Benefits of a Global Overlay Network. *MIT CSAIL TR 2005-070*, December 2005. http://hdl.handle.net/1721.1/30580.
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.