A Comparison of Overlay Routing and Multihoming Route Control[1]

Aditya Akella[2] Jeffrey Pang[2] Anees Shaikh[3]
Bruce Maggs[2] Srinivasan Seshan[2]

August, 2004
CMU-CS-04-158

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

The limitations of BGP routing in the Internet are often blamed for poor end-to-end performance and prolonged connectivity interruptions. Recent work advocates using overlays to effectively bypass BGP's path selection in order to improve performance and fault tolerance. In this paper, we explore the possibility that intelligent control of BGP routes, coupled with ISP multihoming, can provide competitive end-to-end performance and reliability. Using extensive measurements of paths between nodes in a large content distribution network, we compare the relative benefits of overlay routing and multihoming route control in terms of round-trip latency, TCP connection throughput, and path availability. We observe that the performance achieved by route control together with multihoming to three ISPs (3-multihoming), is within 3–12% of overlay routing employed in conjunction 3-multihoming, in terms of both end-to-end RTT and throughput. We also show that while multihoming cannot offer the nearly perfect resilience of overlays, it can eliminate almost all failures experienced by a singly-homed end-network. Our results demonstrate that, by leveraging the capability of multihoming route control, it is not necessary to circumvent BGP routing to extract good wide-area performance and availability from the existing routing system.

[1]This work was supported by the Army Research Office under grant number DAAD19-02-1-0389. Additional support was provided by IBM. A preliminary version of this paper appeared in *Proc. ACM SIGCOMM 2004*, Aug 31 – Sep 3 2004, Portland, OR, USA.
[2]Computer Science Department, Carnegie Mellon University. email: {aditya,jeffpang,srini+,bmm}@cs.cmu.edu
[3]Network Services and Software, IBM T.J. Watson Research Center. email: {aashaikh}@watson.ibm.com

Keywords: performance, availability, multihoming route control, overlay routing

1 Introduction

The limitations of conventional Internet routing based on the Border Gateway Protocol (BGP) are often held responsible for failures and poor performance of end-to-end transfers. A number of studies have shown that the underlying connectivity of the Internet is capable of providing much greater performance and resilience than end-points currently receive. Such studies, exemplified by Detour [25, 26] and RON [6], demonstrate that using *overlay routing* to bypass BGP's policy-driven routing enables quicker reaction to failures and improved end-to-end performance. In this paper, we question whether overlay routing is *required* to make the most of the underlying connectivity, or whether better selection of BGP routes at an end-point is sufficient.

There are two key factors contributing to the differences between overlay routing and BGP-based routing that have not been carefully evaluated in past work: the number of routing choices available to each system and the policies used to select among these routes.

Route Availability. By allowing sources to specify a set of intermediate hops, overlay routing allows end-points nearly arbitrary control over the wide-area path that packets take. On the other hand, BGP only allows a network to announce routes that it actually uses. Thus, to reach a given destination, an end-point has access to only a single path from each Internet Service Provider (ISP) to which it is attached [30]. As a result, an end-point's ability to control routing is tightly linked to the number of ISP connections it has.

Past studies showing the relative benefits of overlay routing draw conclusions based on the highly restrictive case wherein paths from just a single ISP are available [6, 25]. In contrast, in this paper, we carefully consider the degree of ISP multihoming at the end-point, and whether it provides sufficient (BGP) route choices for the end-point to obtain the same performance as when employing an overlay network.

Route Selection. In addition to having a greater selection of routes to choose from than BGP, overlay routing systems use much more sophisticated policies in choosing the route for any particular transfer. Overlays choose routes that optimize end-to-end performance metrics, such as latency. On the other hand, BGP employs much simpler heuristics to select routes, such as minimizing AS hop count or cost. However, this route selection policy is not intrinsic to BGP-based routing – given an adequate selection of BGP routes, end-points can choose the one that results in the best performance, availability, or cost. Several commercial vendors already enable such route control or selection (e.g., [19, 21, 24]).

In this paper, we compare overlays with end-point based mechanisms that use this form of "intelligent" route control of the BGP paths provided by their ISPs. Hereafter, we refer to this as multihoming route control or simply, route control. Notice that we do not assume any changes or improvements to the underlying BGP protocol. Multihoming route control simply allows a multihomed end-network to intelligently schedule its transfers over multiple ISP links in order to optimize performance, availability, cost or a combination of these metrics.

Our goal is to answer the question: *How much benefit does overlay routing provide over BGP, when multihoming and route control are considered?* If the benefit is small, then BGP path selection is not as inferior as it is held to be, and good end-to-end performance and reliability are achievable even when operating completely within standard Internet routing. On the other hand, if overlays yield significantly better performance and reliability characteristics, we have further confirmation of the claim that BGP is fundamentally limited. Then, it is crucial to develop alternate bypass architectures such as overlay routing.

Using extensive active downloads and traceroutes between 68 servers belonging to a large content distribution network (CDN), we compare multihoming route control and overlay routing in terms of three key metrics: round-trip delay, throughput, and availability. Our results suggest that when route control is employed along with multihoming, it can offer performance similar to overlays in terms of round-trip delay and throughput. On average, the round-trip times achieved by the best BGP paths (selected by an ideal route control mechanism using 3 ISPs) are within 3–12% of the best overlay paths (selected by an ideal overlay routing scheme also multihomed to 3 ISPs). Similarly, the throughput on the best overlay paths is only 1–10% better than the best BGP paths. We also show that the marginal difference in the RTT performance can be attributed mainly to overlay routing's ability to select shorter paths, and that this difference can be reduced further if ISPs implement cooperative peering policies. In comparing the end-to-end path availability provided by either approach, we show that multihoming route control, like overlay routing, is able to significantly improve the availability of end-to-end paths.

This paper is structured as follows. In Section 2, we describe past work that demonstrates limitations in the current routing system, including work on overlay routing and ISP multihoming. Section 3 provides an overview of our approach. Section 4 gives details of our measurement testbed. In Section 5, we analyze the RTT and throughput performance differences between route control and overlay routing and consider some possible reasons for the differences. In Section 6, we contrast the end-to-end availability offered by the two schemes. Section 7 discusses the implications of our results and presents some limitations of our approach. Finally, Section 8 summarizes the contributions of the paper.

2 Related Work

Past studies have identified and analyzed several shortcomings in the design and operation of BGP, including route convergence behavior [16, 17] and "inflation" of end-to-end paths due to BGP policies [28, 32]. Particularly relevant to our study are proposals for overlay systems to bypass BGP routing to improve performance and fault tolerance, such as Detour [25] and RON [6].

In the Detour work, Savage et al. [25] study the inefficiencies of wide-area routing on end-to-end performance in terms of round-trip time, loss rate, and throughput. Using observations drawn from active measurements between public traceroute server nodes, they compare the performance on default Internet (BGP) paths with the potential performance from using alternate paths. This work shows that for a large fraction of default paths measured, there are alternate indirect paths offering much better performance.

Andersen *et al.* propose Resilient Overlay Networks (RONs) to address the problems with BGP's fault recovery times, which have been shown to be on order of tens of minutes in some cases [6]. RON nodes regularly monitor the quality and availability of paths to each other, and use this information to dynamically select direct or indirect end-to-end paths. RON mechanisms are shown to significantly improve the availability and performance of end-to-end paths between the overlay nodes. The premise of the Detour and RON studies is that BGP-based route selection is fundamentally limited in its ability to improve performance and react quickly to path failures. Both Detour and RON compare the performance and resilience of overlay paths against default paths via a *single* provider. Overlays offer a greater choice of end-to-end routes, as well as greater flexibility in controlling the route selection. In contrast, we explore the effectiveness of empowering BGP-based route selection with intelligent route control at multihomed end-networks in improving end-to-end availability and performance relative to overlay routing.

Also, several past studies have focused on "performance-aware" routing, albeit not from an end-to-end perspective. Proposals have been made for load sensitive routing within ISPs (see [27], for example) and, intra- and inter-domain traffic engineering [10, 23, 15]. However, the focus of these studies is on balancing the utilization on ISP links and not necessarily on end-to-end performance. More directly related to our work is a recent study on the potential of multihoming route control to improve end-to-end performance and resilience, relative to using paths through a single ISP [3]. Finally, a number of vendors have recently developed intelligent routing appliances that monitor availability and performance over multiple ISP links, and automatically switch traffic to the best provider. These products facilitate very fine-grained selection of end-to-end multihoming routes (e.g., [8, 19, 21, 24]).

3 Comparing BGP Paths with Overlay Routing

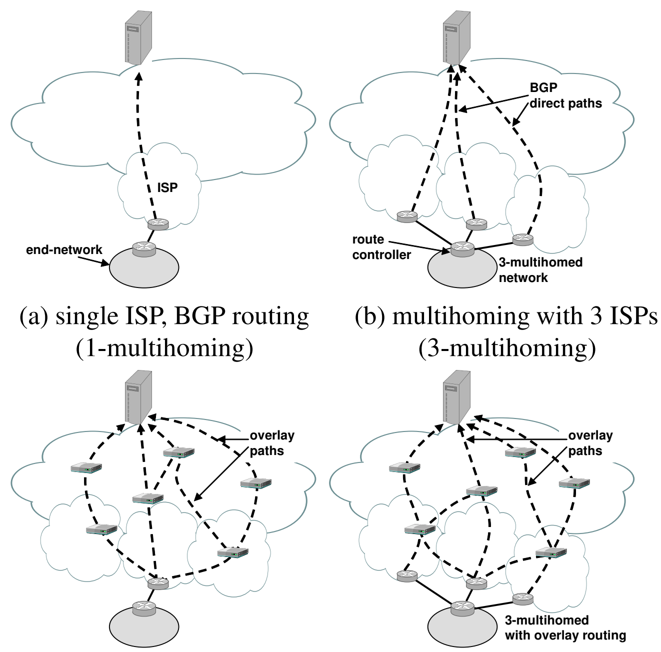
Our objective is to understand whether the modest flexibility of multihoming, coupled with route control, is able to offer end-to-end performance and resilience similar to overlay routing. In order to answer this question, we evaluate an idealized form of multihoming route control where the end-network has instantaneous knowledge about the performance and availability of routes via each of its ISPs for any transfer. We also assume that the end-network can switch between candidate paths to any destination as often as desired. Finally, we assume that the end-network can easily control the ISP link traversed by packets destined for its network (referred to as "inbound control").

In a real implementation of multihoming route control, however, there are practical limitations on the ability of an end-network to track ISP performance, on the rate at which it can switch paths, and on the extent of control over incoming packets. However, recent work [4] shows that simple active and passive measurement-based schemes can be employed to obtain near-optimal availability, and RTT performance that is within 5-10% of the optimal, in practical multihomed environments. Also, simple NAT-based techniques can be employed to achieve inbound route control [4].

To ensure a fair comparison, we study a similarly agile form of overlay routing where the end-point has timely and accurate knowledge of the best performing, or most available, end-to-end overlay paths. Frequent active probing of each overlay link, makes it possible to select and switch to the best overlay path at almost any instant when the size of the overlay network is small (∼50 nodes)[1].

We compare overlay routing and route control with respect to the degree of flexibility available at the end-network. In general, this flexibility is represented by k, the number of ISPs available to either technique at the end-network. In the case of route control, we consider *k-multihoming*, where we evaluate the performance and reliability of end-to-end candidate paths induced by a combination of k ISPs. For overlay routing, we introduce the notion of *k-overlays*, where k is the number of providers available to an end-point for any end-to-end overlay path. In other words, this is simply overlay routing in the presence of k ISP connections.

When comparing k-multihoming with k-overlays, we report results based on the combination of k ISPs that gives the *best performance* (RTT or throughput) across all destinations. In practice an end-network cannot purchase connectivity from all available providers, or easily know which combination of ISPs will provide the best performance. Rather, our results demonstrate how much flexibility is necessary, in terms of the number of ISP connections, and the maximum benefit afforded by this flexibility.



(a) single ISP, BGP routing (1-multihoming)

(b) multihoming with 3 ISPs (3-multihoming)

(c) single ISP, overlay routing (1-overlay)

(d) overlay routing with multihoming (3-overlay)

Figure 1: **Routing configurations:** Figures (a) and (b) show 1-multihoming and 3-multihoming, respectively. Corresponding overlay configurations are shown in (c) and (d), respectively.

Figure 1 illustrates some possible route control and overlay configurations. For example, (a) shows the case of conventional BGP routing with a single default provider (i.e., 1-multihoming). Figure 1(b) depicts end-point route control with three ISPs (i.e., 3-multihoming). Overlay routing with a single first-hop provider (i.e., 1-overlay) is shown in Figure 1(c), and Figure 1(d) shows the case of additional first-hop flexibility in a 3-overlay routing configuration.

We seek to answer the following key questions:

1. On what fraction of end-to-end paths does overlay routing outperform multihoming route control in terms of RTT and throughput? In these cases, what is the extent of the performance difference?

2. What are the reasons for the performance differences? For example, must overlay paths violate inter-domain routing policies to achieve good end-to-end performance?

3. Does route control, when supplied with sufficient flexibility in the number of ISPs, achieve path availability rates that are comparable with overlay routing?

[1] Such frequent probing is infeasible for larger overlays [6].

4 Measurement Testbed

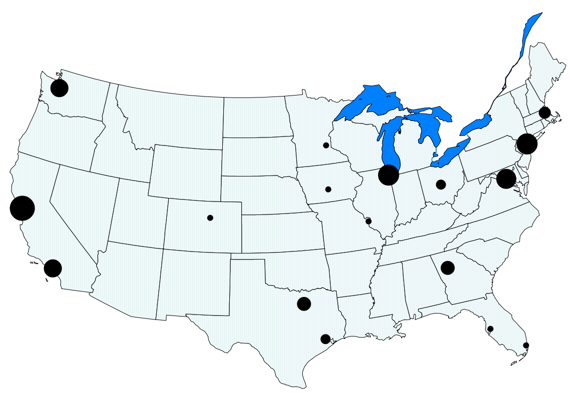
Addressing the questions posed in Section 3 from the perspective of an end-network requires an infrastructure which provides access to a number of BGP path choices via multihomed connectivity, and the ability to select among those paths at a fine granularity. We also require an overlay network with a reasonably wide deployment to provide a good choice of arbitrary wide-area end-to-end paths which could potentially bypass BGP policies.

We address both requirements with a single measurement testbed consisting of nodes belonging to the server infrastructure of the Akamai CDN. Following a similar methodology to that described in [3], we emulate a multihoming scenario by selecting a few nodes in a metropolitan area, each singly-homed to a different ISP, and use them collectively as a stand-in for a multihomed network. Relative to previous overlay routing studies [25, 6], our testbed is larger with 68 nodes. Also, since the nodes are all connected to commercial ISPs, they avoid paths that traverse Internet2, which may introduce unwanted bias due their higher bandwidth and lower likelihood of queuing, compared to typical Internet paths. Our measurements are confined to nodes located in the U.S., though we do sample paths traversing ISPs at all levels of the Internet hierarchy from vantage points in many major U.S. metropolitan areas.

The 68 nodes in our testbed span 17 U.S. cities, averaging about four nodes per city, connected to commercial ISPs of various sizes. The nodes are chosen to avoid multiple servers attached to the same provider in a given city. The list of cities and the tiers of the corresponding ISPs are shown in Figure 2(a). The tiers of the ISPs are derived from the work in [31]. The geographic distribution of the testbed nodes is illustrated in Figure 2(b). We emulate multihomed networks in 9 of the 17 metropolitan areas where there are at least 3 providers – Atlanta, Bay Area, Boston, Chicago, Dallas, Los Angeles, New York, Seattle and Washington D.C.

City	**Providers/tier**				
	1	**2**	**3**	**4**	**5**
Atlanta, GA	2	0	1	1	0
Bay Area, CA	5	0	3	1	2
Boston, MA	1	0	1	0	1
Chicago, IL	6	1	0	1	0
Columbus, OH	0	1	0	1	0
Dallas, TX	3	0	0	1	0
Denver, CO	1	0	0	0	0
Des Moines, IO	0	1	0	0	0
Houston, TX	1	1	0	0	0
Los Angeles, CA	3	0	3	0	0
Miami, FL	1	0	0	0	0
Minneapolis, MN	0	0	1	0	0
New York, NY	3	2	2	1	0
Seattle, WA	2	0	2	1	1
St Louis, MO	1	0	0	0	0
Tampa, FL	0	1	0	0	0
Washington DC	3	0	3	0	2

(a) Testbed ISPs



(b) Node locations

Figure 2: **Testbed details:** The cities and distribution of ISP tiers in our measurement testbed are listed in (a). The geographic location is shown in (b). The area of each dot is proportional to the number of nodes in the region.

5 Latency and Throughput Performance

We now present our results on the relative latency and throughput performance benefits of multihoming route control compared with overlay routing. We first describe our data collection methodology (Section 5.1) and evaluation metrics (Section 5.2). Then, we present the key results in the following order. First we compare 1-multihoming against 1-overlays along the same lines as the analysis in [25] (Section 5.3). Next, we compare the benefits of using k-multihoming and k-overlay routing, relative to using default paths through a single provider (Section 5.4). Then, we compare k-multihoming against 1-overlay routing, for $k \geq 1$ (Section 5.5). Here, we wish to quantify the benefit to end-systems of greater flexibility in the choice of BGP routes via multihoming, relative to the power of 1-overlays. Next, we contrast k-multihoming against k-overlay routing to understand the additional benefits gained by allowing end-systems almost arbitrary control on end-to-end paths, relative to multihoming (Section 5.6). Finally, we examine some of the underlying reasons for the performance differences (Sections 5.7 and 5.8).

5.1 Data Collection

Our comparison of overlays and multihoming is based on observations drawn from two data sets collected on our testbed. The first data set consists of active HTTP downloads of small objects (10 KB) to measure the *turnaround times* between the pairs of nodes. The turnaround time is the time between the transfer of the last byte of the HTTP request and the receipt of the first byte of the response, and provides an estimate of the round-trip time. Hereafter, we will use the terms turnaround time and round-trip time interchangeably. Every 6 minutes, turnaround time samples are collected between all node-pairs (including those within the same city).

The second data set contains "throughput" measurements from active downloads of 1 MB objects between the same set of node-pairs. These downloads occur every 30 minutes between all node-pairs. Here, throughput is simply the size of the transfer (1 MB) divided by the time between the receipt of the first and last bytes of the response data from the server (source). As we discuss in Section 5.2, this may not reflect the steady-state TCP throughput along the path.

Since our testbed nodes are part of a production infrastructure, we limit the frequencies at which all-pairs measurements are collected as described above. To ensure that all active probes between pairs of nodes observe similar network conditions, we scheduled them to occur within a 30 second interval for the round-trip time data set, and within a 2 minute interval for the throughput data set. For the latter, we also ensure that an individual node is involved in at most one transfer at any time so that our probes do not contend for bandwidth at the source or destination network. The transfers may interfere elsewhere in the Internet, however. Also, since our testbed nodes are all located in the U.S., the routes we probe, and consequently, our observations, are U.S.-centric.

The round-trip time data set was collected from Thursday, December 4th, 2003 through Wednesday, December 10th, 2003. The throughput measurements were collected between Thursday, May 6th, 2004 and Tuesday, May 11th, 2004 (both days inclusive).

5.2 Performance Metrics

We compare overlay routing and multihoming according to two metrics derived from the data above: round-trip time (RTT) and throughput. In the RTT data set, for each 6 minute measurement interval, we build a weighted graph over all the 68 nodes where the edge weights are the RTTs measured between the corresponding node-pairs. We then use Floyd's algorithm to compute the shortest paths between all node-pairs. We estimate the RTT performance from using k-multihoming to a given destination by computing the minimum of the RTT estimates along the direct paths from the k ISPs in a city to the destination node (i.e., the RTT measurements between the Akamai CDN nodes representing the k ISPs and the destination node). To estimate the performance of k-overlay routing, we compute the shortest paths from the k ISPs to the destination node and choose the minimum of the RTTs of these paths.

Note that we do not prune the direct overlay edge in the graph before performing the shortest path computation. As a result, the shortest overlay path between two nodes could be a *direct* path (i.e., chosen by BGP). Hence our comparison is not limited to direct versus indirect paths, but is rather between direct and *overlay* paths. In contrast, the comparison in [25] is between the direct path and the *best indirect path*.

For throughput, we similarly construct a weighted, directed graph between all overlay nodes every 30 minutes (i.e., our 1 MB object download frequency). The edge weights are the throughputs of the 1 MB transfers (where throughput is computed as described in Section 5.1). We compute the throughput performance of k-multihoming and k-overlay routing similar to the RTT performance computation above. Notice, however, that computing the overlay throughput performance is non-trivial and is complicated by the problem of estimating the end-to-end throughput for a 1 MB TCP transfer on indirect overlay paths.

Our approach here is to use round-trip time and throughput measurements on individual overlay hops to first compute the underlying loss rates. Since it is likely that the paths we measure do not observe any loss, thus causing the transfers to likely remain in their slow-start phases, we use the small connection latency model developed in [7]. The typical MSS in our 1MB transfers is 1460 bytes. Also, the initial congestion window size is 2 segments and there is no initial 200ms delayed ACK timeout on the first transfer. In the throughput data set, we measure a mean loss rate of 1.2% and median, 90th, 95th and 99th percentile loss rates of 0.004%, 0.5%, 1% and 40% across all paths measured, respectively.

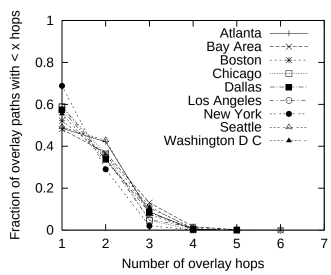
We can then use the sum of round-trip times and a combination of loss rates on the individual hops as the end-to-end round-trip time and loss rate estimates, respectively, and employ the model in [7] to compute the end-to-end overlay throughput for the 1 MB transfers. To combine loss rates on individual links, we follow the same approach as that described in [25]. We consider two possible combination functions. The first, called *optimistic*, uses the maximum observed loss on any individual overlay hop along an overlay path as an estimate of the end-to-end overlay loss rate. This assumes that the TCP sender is primarily responsible for the observed losses. In the *pessimistic* combination, we compute the end-to-end loss rate as the sum of individual overlay hop loss rates, assuming the losses on each link to be due to independent background traffic in the network[2]. Due to the complexity of computing arbitrary length throughput-maximizing overlay paths, we only consider indirect paths comprised of at most two overlay hops in our throughput comparison.

5.3 1-Multihoming versus 1-Overlays

First, we compare the performance of overlay routing against default routes via a single ISP (i.e., 1-overlay against 1-multihoming), along the same lines as [25]. Note that, in the case of 1-overlays, the overlay path from a source node may traverse through any intermediate node, including nodes located in the same city as the source.

City	1-multihoming/ 1-overlay
Atlanta	1.35
Bay Area	1.20
Boston	1.28
Chicago	1.29
Dallas	1.32
Los Angeles	1.22
New York	1.29
Seattle	1.34
Wash D.C.	1.30
Average	1.29

(a) 1-multihoming RTT relative to 1-overlays



(b) 1-overlay path length

Figure 3: **Round-trip time performance:** Average RTT performance of 1-multihoming relative to 1-overlay routing is tabulated in (a) for various cities. The graph in (b) shows the distribution of the number of overlay hops in the best 1-overlay paths, which could be the direct path (i.e., 1 overlay hop).

Round-trip time performance. Figure 3(a) shows the RTT performance of 1-multihoming relative to 1-overlay routing. Here, the performance metric (y-axis) reflects the relative RTT from 1-multihoming versus the RTT when using 1-overlays, averaged over all samples to all destinations. The difference between this metric and 1 represents the relative advantage of 1-overlay routing over 1-multihoming. Notice also that since the best overlay path could be the direct BGP path, the performance from overlays is at least as good as that from the direct BGP path. We see from the table that overlay routing can improve RTTs between 20% and 35% compared to using direct BGP routes over a single ISP. The average improvement is about 29%. The observations in [25] are similar.

We show the distribution of overlay path lengths in Figure 3(b), where the direct (BGP) path corresponds to a single overlay hop. Notice that in most cities, the best overlay path is only one or two hops in more than 90% of the measurements. That is, the majority of the RTT performance gains in overlay networks are realized without requiring more than a single intermediate hop. Also, on an average, the best path from 1-overlays coincides with the direct BGP path in about 55% of the measurements (average y-axis value at x=1 across all cities).

Throughput performance. In Table 1, we show the throughput performance of 1-overlays relative to 1-multihoming for both the pessimistic and the optimistic estimates. 1-overlays achieve 6–20% higher throughput than 1-multihoming, according to the pessimistic estimate. According to the optimistic throughput estimate, 1-overlays achieve 10–25% better throughput. In Table 1, we also show the fraction of times an indirect overlay path obtains better throughput than the direct path, for either throughput estimation function. Under the pessimistic throughput estimate, on average,

[2]The end-to-end loss rate over two overlay links with independent loss rates of p_1 and p_2 is $1-(1-p_1)(1-p_2) = p_1+p_2-p_1p_2$. p_1p_2 is negligible in our measurements, so we ignore it.

City	Pessimistic estimate		Optimistic estimate	
	Throughput metric	*Fraction of indirect paths*	*Throughput metric*	*Fraction of indirect paths*
Atlanta	1.14	17%	1.17	21%
Bay Area	1.06	11%	1.10	22%
Boston	1.19	22%	1.24	26%
Chicago	1.12	13%	1.15	18%
Dallas	1.16	18%	1.18	22%
Los Angeles	1.18	15%	1.21	17%
New York	1.20	14%	1.25	26%
Seattle	1.18	28%	1.25	35%
Wash D.C.	1.09	13%	1.13	18%
Average	1.15	17%	1.19	23%

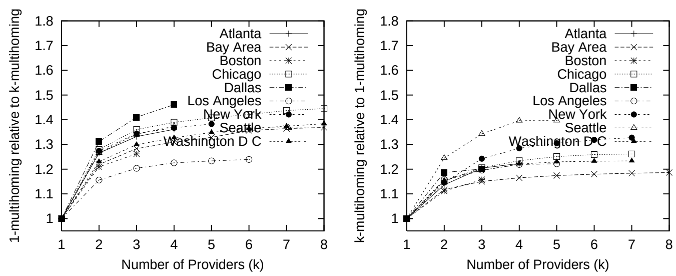
Table 1: **Throughput performance:** This table shows the 1 MB TCP transfer performance of 1-overlay routing relative to 1-multihoming (for both estimation functions). Also shown is the fraction of measurements in which 1-overlay routing selects an indirect path in each city.

1-overlay routing benefits from employing an indirect path in about 17% of the cases. Under the optimistic estimate, this fraction is 23%.

Summary. 1-Overlays offer significantly better round-trip time performance than 1-multihoming (29% on average). The throughput benefits are lower, but still significant (15% on average). Also, in a large fraction of the measurements, indirect 1-overlay paths offer better RTT performance than direct 1-multihoming paths.

5.4 1-Multihoming versus k-Multihoming and k-Overlays

In this section we compare the flexibility offered by multihoming route control at an end point in isolation, and in combination with overlay routing, against using default routes via a single ISP (i.e., k-multihoming and k-overlays against 1-multihoming). The main purpose of these comparisons is to establish a baseline for the upcoming head-to-head comparisons between k-multihoming and k-overlay routing in Sections 5.5 and 5.6.

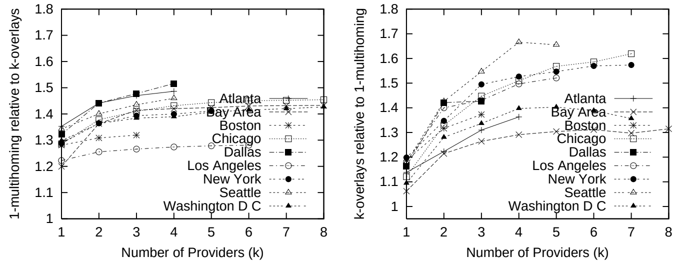


(a) Relative RTTs (b) Throughput (pessimistic)

Figure 4: **Benefits of k-multihoming:** The RTT of 1-multihoming relative to k-multihoming is shown in (a) and throughput (pessimistic estimate) of k-multihoming relative to 1-multihoming is shown in (b).

1-Multihoming versus k-multihoming. Figure 4(a) shows the RTT performance of 1-multihoming relative to the RTT performance from k-multihoming averaged across all samples to all destinations (y-axis), as a function of the number of providers, k (x-axis). Note that the difference between the performance metric on the y-axis and 1 indicates the relative advantage of k-multihoming over 1-multihoming. The RTT benefit from multihoming is about 15–30% for $k = 2$ and about 20–40% for $k = 3$ across all the cities. Also, beyond $k = 3$ or 4 the marginal improvement in the RTT performance from multihoming is negligible. The observations made by Akella et al. in [3] are similar.

Figure 4(b) similarly shows the throughput performance of k-multihoming relative to the throughput from 1-multihoming, according to the pessimistic estimate. The results for the optimistic estimate are similar and are omitted for brevity. Again, k-multihoming, for $k = 3$, achieves 15–25% better throughput than 1-multihoming and the marginal improvement in the throughput performance is negligible beyond $k = 3$.



(a) Relative RTTs (b) Throughput (pessimistic)

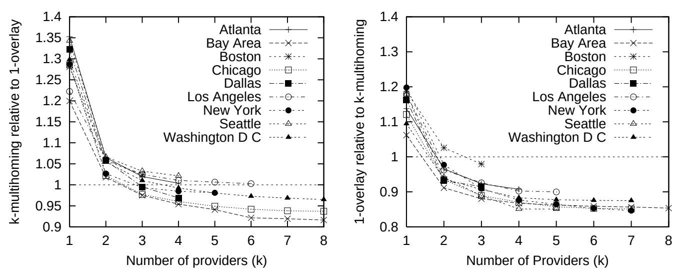
Figure 5: **Benefits of k-overlays:** The RTT of 1-multihoming relative to k-overlays is shown in (a) and throughput (pessimistic estimate) of k-overlays relative to 1-multihoming is shown in (b).

1-Multihoming versus k-overlays. In Figure 5(a), we show the RTT performance of 1-multihoming relative to k-overlays as a function of k. Notice that k-overlay routing achieves 25–50% better RTT performance than 1-multihoming, for $k = 3$. Notice also, that the RTT performance from k-overlay routing, for $k \geq 3$, is about 5–20% better than that from 1-overlay routing. Figure 5(b) similarly compares the throughput performance of k-overlays relative to 1-multihoming, for the pessimistic estimate. Again, 3-overlay routing, for example, is 20–55% better than 1-multihoming and about 10–25% better than 1-overlay routing. The benefit beyond $k = 3$ is marginal across most cities, for both RTT as well as throughput.

Summary. Both k-multihoming and k-overlay routing offer significantly better performance than 1-multihoming, in terms of both RTT and throughput. In addition, k-overlay routing, for $k \geq 3$ achieves significantly better performance compared to 1-overlay routing (5–20% better according to RTT and 10–25% better according to throughput).

5.5 k-Multihoming versus 1-Overlays

So far, we have evaluated multihoming route control (i.e., k-multihoming for $k \geq 2$) and overlay routing in isolation of each other. In what follows, we provide a head-to-head comparison of the two systems. First, in this section, we allow end-points the flexibility of multihoming route control and compare the resulting performance against 1-overlays.



(a) Relative RTTs (b) Throughput (pessimistic)

Figure 6: **Multihoming versus 1-overlays:** The RTT of k-multihoming relative to 1-overlays is shown in (a) and throughput (pessimistic) of 1-overlays relative to k-multihoming in (b).

In Figure 6, we plot the performance of k-multihoming relative to 1-overlay routing. Here, we compute the average ratio of the best RTT or throughput to a particular destination, as achieved by either technique. The average is taken over paths from each city to destinations in other cities, and over time instants for which we have a valid measurement over all ISPs in the city.[3] We also note that in all but three cities, the best 3-multihoming providers according to RTT

[3] Across all cities, an average of 10% of the time instants did not have a valid measurement across all providers; nearly all of these cases were due to limitations in our data collection infrastructure, and not failed download attempts.

were the same as the best 3 according to throughput; in the three cities where this did not hold, the third and fourth best providers were simply switched and the difference in throughput performance between them was less than 3%.

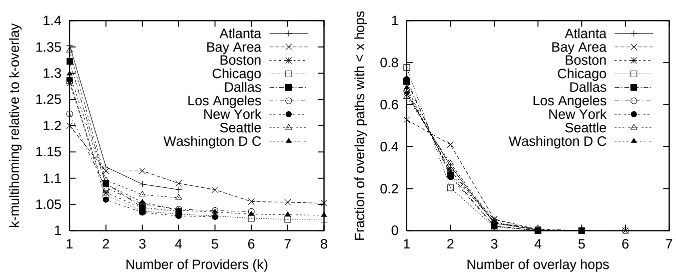
The comparison according to RTT is shown in Figure 6(a). The relative performance advantage of 1-overlays is less than 5% for $k = 3$ in nearly all cities. In fact, in some cities, e.g., Bay Area and Chicago, 3-multihoming is marginally better than overlay routing. As the number of ISPs is increased, multihoming is able to provide shorter round-trip times than overlays. Figure 6(b) shows relative benefits according to the pessimistic throughput estimate. Here, multihoming for $k \geq 3$ actually provides 2–12% better throughput than 1-overlays across all cities. The results are similar for the optimistic computation and are omitted for brevity.

Summary. The performance advantages of 1-overlays are vastly reduced (or eliminated) when the end-point is allowed greater flexibility in the choice of BGP paths via multihoming route control.

5.6 k-Multihoming versus k-Overlays

In the previous section, we evaluated 1-overlay routing, where all overlay paths start from a single ISP in the source city. In this section, we allow overlays additional flexibility by permitting them to initially route through more of the available ISPs in each source city. Specifically, we compare the performance benefits of k-multihoming against k-overlay routing.

In the case of k-overlays, the overlay path originating from a source node may traverse any intermediate nodes, including those located in the same city as the source. Notice that the performance from k-overlays is at least as good as that from k-multihoming (since we allow overlays to take the direct path). The question, then, is how much more advantage do overlays provide if multihoming is already employed by the source.



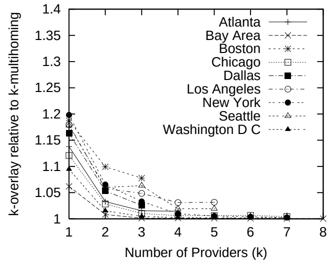
(a) Relative RTTs (b) 3-Overlay path length

Figure 7: **Round-trip time improvement:** Round-trip time from k-multihoming relative to k-overlay routing, as a function of k, is shown in (a). In (b), we show the distribution of the number of overlay hops in the best k-overlay paths, for k=3.

Round-trip time performance. Figure 7(a) shows the improvement in RTT for k-multihoming relative to k-overlays, for various values of k. We see that on average, for $k = 3$, overlays provide 3–12% better RTT performance than the best multihoming solution in most of the cities in our study. The performance gap between multihoming and overlays is less significant for $k \geq 4$.

Figure 7(b) shows the distribution of the number of overlay hops in the paths selected by 3-overlay routing optimized for RTT. The best overlay path coincides with the best 3-multihoming BGP path in 67% of the cases, on average across all cities. Recall that the corresponding fraction for 1-overlay routing in Figure 3(b) was 55%. With more ISPs to links to choose from, overlay routing selects a *higher* fraction of direct BGP paths, as opposed to choosing from the greater number of indirect paths also afforded by multihoming.

Throughput performance. Figure 8(a) shows the throughput performance of k-multihoming relative to k-overlays using the pessimistic throughput estimation function. From this figure, we see that multihoming achieves throughput performance within 1–10% of overlays, for $k = 3$. The performance improves up to $k = 3$ or $k = 4$. In all the cities, the throughput performance of 4-multihoming is within 3% of overlay routing. In Figure 8(b), we also show the fraction of measurements where an indirect 3-overlay path offers better performance than the direct 3-multihoming path, for the pessimistic throughput estimate. On average, this fraction is about 8%. Notice that this is again lower than the corresponding percentage for 1-overlays from Table 1 ($\approx$ 17%).



(a) Throughput improvement (pessimistic estimate)

City	Fraction of indirect paths
Atlanta	5%
Bay Area	1%
Boston	13%
Chicago	3%
Dallas	8%
Los Angeles	4%
New York	8%
Seattle	31%
Wash D.C.	2%
Average	8%

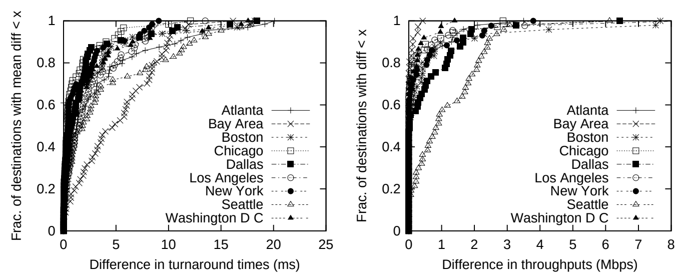
(b) Fraction of indirect paths in 3-overlay routing

Figure 8: **Throughput improvement:** Throughput performance of k-multihoming relative to k-overlays for various cities is shown in (a). The table in (b) shows the fraction of measurements on which k-overlay routing selected an indirect end-to-end path, for the case of $k = 3$.

Summary. When employed in conjunction with multihoming, overlay routing offers marginal benefits over employing multihoming alone. For example, multiple ISPs allows overlay routing to achieve only a 3–12% RTT improvement over multihoming route control (for $k = 3$), and 1–10% improvement in throughput. In addition, k-overlay routing selects a larger fraction of direct BGP-based end-to-end paths, compared to 1-overlay routing.

5.7 Unrolling the Averages

So far, we presented averages of the performance differences for various forms of overlay routing and multihoming route control. In this section, focusing on 3-overlays and 3-multihoming, we present the underlying distributions in the performance differences along the paths we measure. Our goal in this section is to understand if the averages are particularly skewed by: (1) certain destinations, for each source city or (2) a few measurement samples on which overlays offer significantly better performance than multihoming or (3) by time-of-day or day-of-week effects.



(a) Mean difference in round-trip times

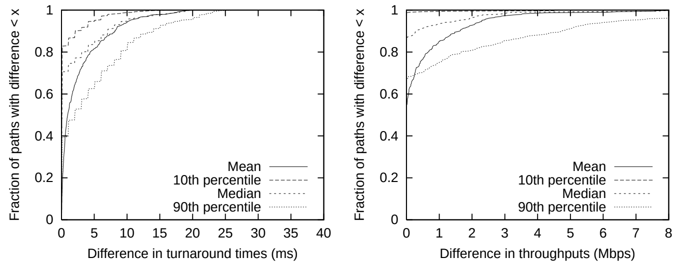
(b) Mean difference in throughputs (pessimistic)

Figure 9: **Performance per destination:** Figure (a) is a CDF of the mean difference in RTTs along the best overlay path and the best direct path, across paths measured from each city. Similarly, Figure (b) plots the CDF of the mean difference in throughputs (pessimistic estimate).

Performance per destination. In Figure 9(a), for each city, we show the distribution of the average difference in RTT between the best 3-multihoming path and the best 3-overlay path to each destination (i.e., each point represents one destination). In most cities, the average RTT differences across 80% of the destinations are less than 10ms. Notice that in most cities, the difference is greater than 15ms for less than 5% of the destinations.

In Figure 9(b), we consider the distribution of the average throughput difference of the best 3-multihoming path and the best 3-overlay path for the pessimistic estimate of throughput. We see the throughput difference is less than 1 Mbps for 60–99% of the destinations. We also note that, for 1–5% of the destinations, the difference is in excess of 4

Mbps. Recall from Figure 8, however, that these differences result in an average relative performance advantage for overlays of less than 1–10% (for $k = 3$).



(a) Round-trip time (b) Throughput (pessimistic)

Figure 10: **Underlying distributions:** Figure showing the mean, median, 10th percentile and 90th percentile difference across various source-destination pairs. Figure (a) plots RTT, while figure (b) plots throughput (pessimistic estimate).

Mean versus other statistics. In Figures 10(a) and (b) we plot the average, median, and 10th and 90th percentiles of the difference in RTT and (pessimistic) throughput, respectively, between the best 3-multihoming option and the best 3-overlay path across paths in all cities. In Figure 10(a) we see that the median RTT difference is fairly small. More than 90% of the median RTT differences are less than 10ms. The 90th percentile of the difference is marginally higher with roughly 10% greater than 15ms. The median throughput differences in Figure 10(b) are also relatively small – less than 500 kbps about 90% of the time. Considering the upper range of the throughput difference (i.e., the 90th percentile difference), we see that a significant fraction (about 20%) are greater than 2 Mbps. These results suggest that the absolute round-trip and throughput differences between multihoming and overlay routing are small for the most part, though there are a few of cases where differences are more significant, particularly for throughput.

Time-of-day and day-of-week effects. We also considered the effects of daily and weekly network usage patterns on the relative performance of k-multihoming and k-overlays. It might be expected that route control would perform worse during peak periods since overlay paths have greater freedom to avoid congested parts of the network. We do not see any discernible time-of-day effects in paths originating from a specific city, however, both in terms of RTT and throughput performance.

Similarly, we also examine weekly patterns to determine whether the differences are greater during particular days of the week, but again there are no significant differences for either RTT or throughput. We omit both these results for brevity. The lack of a time-of-day effect on the relative performance may be indicative that ISP network operators already take such patterns into account when performing traffic engineering.

Summary. k-overlays offer significantly better performance relative to k-multihoming for a small fraction of transfers from a given city. We observed little dependence on the time-of-day or day-of-week in the performance gap between overlays and multihoming.

5.8 Reasons for Performance Differences

Next, we try to identify the underlying causes of performance differences between k-multihoming and k-overlay routing. We focus on the RTT performance and the case where $k = 3$. First, we ask if indirect paths primarily improve propagation delay or mostly select less congested routes than the direct paths. Then, we focus on how often the best-performing indirect paths violate common inter-domain and peering policies.

5.8.1 Propagation Delay and Congestion Improvement

In this section, we are interested in whether the modest advantage we observe for overlay routing is due primarily to its ability to find "shorter" (i.e., lower propagation delay) paths outside of BGP policy routing, or whether the gains come from being able to avoid congestion in the network (a similar analysis was done in [25]).

The pairwise instantaneous RTT measurements we collect may include a queuing delay component in addition to the base propagation delay. When performance improvements are due primarily to routing around congestion, we expect the difference in propagation delay between the indirect and direct path to be small. Similarly, when the propagation difference is large, we can attribute the performance gain to the better efficiency of overlay routing compared to BGP in choosing "shorter" end-to-end paths. In our measurements, to estimate the propagation delay on each path, we take the 5th percentile of the RTT samples for the path.

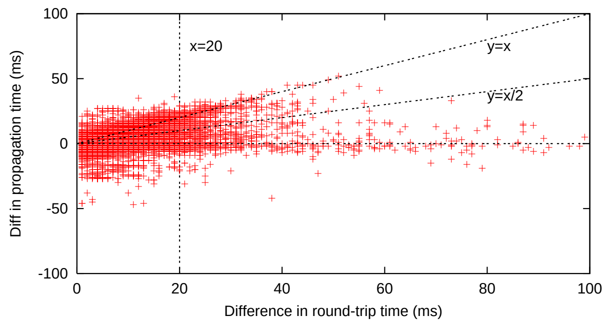


Figure 11: **Propagation vs congestion:** A scatter plot of the RTT improvement (x-axis) vs propagation time improvement (y-axis) of the indirect overlay paths over the direct paths.

In Figure 11, we show a scatter plot of the overall RTT improvement (x-axis) and the corresponding propagation time difference (y-axis) offered by the best overlay path relative to the best multihoming path. The graph only shows measurements in which the indirect overlay paths offer an improved RTT over the best direct path. Points near the $y = 0$ line represent cases in which the RTT improvement has very little associated difference in propagation delay. Points near the $y = x$ line are paths in which the RTT improvement is primarily due to better propagation time.

For paths with a large RTT improvement (e.g., > 50ms), the points are clustered closer to the $y = 0$ line, suggesting that large improvements are due primarily to routing around congestion. We also found, however, that 66% of all the points lie above the $y = x/2$ line. These are closer to the $y = x$ line than $y = 0$, indicating that a majority of the round-trip improvements do arise from a reduction in propagation delay. In contrast, Savage et al. [25] observe that both avoiding congestion and the ability to find shorter paths are equally responsible for the overall improvements from overlay routing. The difference in our observations from those in [25] could be due to the fact that Internet paths are better provisioned and less congested today than 3-4 years ago. However, they are sometimes circuitous, contributing to inflation in end-to-end paths [28].

Total fraction of lower delay overlay paths	33%	
	Fraction of lower delay paths	Fraction of all overlay paths
Indirect paths with $>$ 20ms improvement	4.8%	1.6%
Prop delay improvement $<$ x% of overall improvement (whenever overall improvement $>$ 20ms)		
$<$ 50%	2.2%	0.7%
$<$ 25%	1.7%	0.6%
$<$ 10%	1.3%	0.4%

Table 2: **Analysis of overlay paths:** Classification of indirect paths offering $>$ 20ms improvement in RTT performance.

To further investigate the relative contributions of propagation delay and congestion improvements, we focus more closely on cases where indirect overlay paths offer a significant improvement ($>$ 20ms) over the best direct paths. Visually, these are all points lying to the right of the $x = 20$ line in Figure 11. In Table 2 we present a classification of all of the indirect overlay paths offering $>$ 20ms RTT improvement. Recall that, in our measurement, 33% of the indirect 3-overlay paths had a lower RTT than the corresponding best direct path (Section 5.6, Figure 7 (b)). However, of these paths, only 4.8% improved the delay by more than 20ms (Table 2, row 3). For less than half of these, or 2.2% of all lower delay overlay paths, the propagation delay improvement relative to direct paths was less than 50% of the overall RTT improvement. Visually, these points lie to the right of $x = 20$ and below the $y = x/2$ lines in Figure 11. Therefore, these are paths where the significant improvement in performance comes mainly from the ability of the overlay to avoid congested links. Also, when viewed in terms of all overlay paths (see Table 2, column 3), we see that these paths form a very small fraction of all overlay paths ($\approx 0.7\%$).

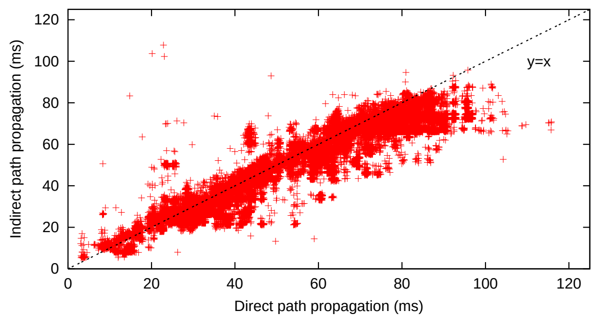


Figure 12: **"Circuitousness" of routes:** Figure plotting the propagation delay of the best indirect path (y-axis) against the best multihoming path (x-axis).

Finally, if we consider the propagation delay of the best indirect overlay path versus the best multihoming path, we can get some idea of the relative ability to avoid overly "circuitous" paths, arising from policy routing, for example. Figure 12 shows a scatter plot of the propagation delay of the best direct path from a city (x-axis) and the best propagation delay via an indirect path (y-axis). Again, points below the $y = x$ line are cases in which overlay routing finds shorter paths than conventional BGP routing, and vice versa. Consistent with the earlier results, we see that the majority of points lie below the $y = x$ line where overlays find lower propagation delay paths. Moreover, for cases in which the direct path is shorter (above the $y = x$ line), the difference is generally small, roughly 10-15ms along most of the range.

Summary. A vast majority of RTT performance improvements from overlay routing arise from its ability to find shorter end-to-end paths compared to the best direct BGP paths. However, the most significant improvements ($>$ 50ms) stem from the ability of overlay routing to avoid congested ISP links[4].

5.8.2 Inter-domain and Peering Policy Compliance

To further understand the performance gap between some overlay routes and direct BGP routes, we categorize the overlay routes by their compliance with common inter-domain and peering policies. Inter-domain and peering policies typically represent business arrangements between ISPs [11, 20]. Because end-to-end overlay paths need not adhere to such policies, we try to quantify the performance gain that can be attributed to ignoring them.

Two key inter-domain policies [12] are *valley-free routing* — ISPs generally do not provide transit between their providers or peers because it represents a cost to them; and *prefer customer* — when possible, it is economically preferable for an ISP to route traffic via customers rather than providers or peers, and peers rather than providers. In addition, Spring et al. [28] observed that ISPs often obey certain *peering policies*. Two common policies are *early exit* — in which ISPs "offload" traffic to peers quickly by using the peering point closest to the source; and *late exit* — some ISPs cooperatively carry traffic further than they have to by using peering points closer to the destination. BGP path selection is also impacted by the fact that the routes must have the shortest AS hop count.

We focus on indirect overlay paths (i.e., > 1 virtual hop) that provide better end-to-end *round-trip* performance than the corresponding direct BGP paths. To characterize these routes, we identified AS level paths using traceroutes performed during the same period as the turnaround time measurements. Each turnaround time measurement was matched with a traceroute that occurred within 20 minutes of it (2.7% did not have corresponding traceroutes and were ignored in this analysis). We map IP addresses in the traceroute data to AS numbers using a commercial tool which uses BGP tables from multiple vantage points to extract the "origin AS" for each IP prefix [2]. One issue with

[4]The improvements from overlay routing could also be from overlays choosing higher bandwidth paths. This aspect is difficult to quantify and we leave it as future work.

deriving the AS path from traceroutes is that these router-level AS paths may be different than the actual BGP AS path [18, 5, 14], often due to the appearance of an extra AS number corresponding to an Internet exchange point or a sibling AS[5]. In our analysis, we omit exchange point ASes, and also combine the sibling ASes, for those that we are able to identify. To ascertain the policy compliance of the indirect overlay paths, we used AS relationships generated by the authors of [31] during the same period as our measurements.

In our AS-level overlay path construction, we ignore the ASes of intermediate overlay nodes if they were used merely as non-transit hops to connect overlay path segments. For example, consider the overlay path between a source in AS `S1` and a destination in `D2`, composed of the two AS-level segments `S1 A1 B1 C1` and `C1 B2 D2`, where the intermediate node is located in `C1`. If the time spent in `C1` is short ($<$ 3ms), and `B1` and `B2` are the same ISP, we consider the AS path as `S1 A1 B1 D2`, otherwise we consider it as `S1 A1 B1 C1 B2 D2`. Since we do this only for intermediate ASes that are not a significant factor in the end-to-end round-trip difference, we avoid penalizing overlay paths for policy violations that are just artifacts of where the intermediate hop belongs in the AS hierarchy.

	Improved Overlay Paths			>20ms Imprv Paths		
	%	*RTT Imprv (ms)*		%	*RTT Imprv (ms)*	
		Avg	90th		Avg	90th
Violates Inter-Domain Policy	**66.8**	8.3	17	**68.7**	33.7	40
Valley-Free Routing	61.0	8.2	17	58.5	33.7	40
Prefer Customer	14.9	8.9	18	16.3	41.3	47
Valid Inter-Domain Path	**25.2**	7.3	15	**19.4**	36.1	44
Same AS-Level Path	15.3	6.9	13	9.4	40.9	53
Earlier AS Exit	1.9	5.6	10	0.8	43.2	51
Similar AS Exits	6.9	6.4	12	4.9	39.6	55
Later AS Exit	6.5	7.9	14	3.7	42.1	51
Diff AS-Level Path	9.9	8.0	17	10.0	31.5	39
Longer than BGP Path	4.5	7.6	17	4.6	30.9	43
Same Len as BGP Path	4.8	8.6	18	5.3	32.0	37
Shorter than BGP Path	0.6	6.2	9	0.1	36.4	55
Unknown	**8.0**			**11.9**		

Table 3: **Overlay routing policy compliance:** Breakdown of the mean and 90th percentile round trip time improvement of indirect overlay routes by: (1) routes did not conform to common inter-domain policies, and (2) routes that were valid inter-domain paths but either exited ASes at different points than the direct BGP route or were different than the BGP route.

Table 3 classifies the indirect overlay paths by policy conformance. As expected, the majority of indirect paths (67%) violated either the valley-free routing or prefer customer policies. However, a large fraction of overlay paths (25%) appeared to be policy compliant. We sub-categorize the latter fraction of paths further by examining which AS-level overlay paths were identical to the AS-level direct BGP path and which ones were different.

For each overlay path that was identical, we characterized it as exiting an AS earlier than the direct path if it remained in the AS for at least 20ms less than it did in the direct path. We characterized it as exiting later if it remained in an AS for at least 20ms longer. We consider the rest of the indirect paths to be “similar” to the direct BGP paths. We see that almost all identical AS-level overlay paths either exited later or were similar to the direct BGP path. This suggests that cooperation among ISPs, e.g., in terms of late exit policies, can improve performance on BGP routes and further close the gap between multihoming and overlays. We also note that for the AS-level overlay paths that differed, the majority were the same length as the corresponding direct path chosen by BGP.

Summary. In achieving better RTT performance than direct BGP paths, most indirect overlay paths violate common inter-domain routing policies. We observed that a fraction of the policy-compliant overlay paths could be realized by BGP if ISPs employed cooperative peering policies such as late exit.

6 Resilience to Path Failures

BGP’s policy-based routing architecture masks a great deal of topology and path availability information from end-networks in order to respect commercial relationships and limit the impact of local changes on neighboring down-

[5]Two ASes identified as peers may actually be siblings [31, 11], in which case they would provide transit for each other’s traffic because they are administered by the same entity. We classified peers as siblings if they appeared to provide transit in the direct BGP paths in our traceroutes, and also manually adjusted pairings that were not related.

stream ASes [10, 22]. This design, while having advantages, can adversely affect the ability of end-networks to react quickly to service interruptions since notifications via BGP's standard mechanisms can be delayed by tens of minutes [16]. Networks employing multihoming route control can mitigate this problem by monitoring paths across ISP links, and switching to an alternate ISP when failures occur. Overlay networks provide the ability to quickly detect and route around failures by frequently probing the paths between all overlay nodes.

In this section, we perform two separate, preliminary analyses to assess the ability of both mechanisms to withstand end-to-end path failures and improve availability of Internet paths. The first approach evaluates the availability provided by route control based on active probe measurements on our testbed. In the second we compute the end-to-end path availability from both route control and overlays using estimated availabilities of routers along the paths.

6.1 Active Measurements of Path Availability

In our first approach, we perform two-way ICMP pings between the 68 nodes in our testbed. The ping samples were collected between all node-pairs over a five day period from January 23rd, 2004 to January 28th, 2004. The probes are sent once every minute with a one second timeout. If no response is received within a second, the ping is deemed lost. A path is considered to have failed if ≥ 3 consecutive pings (each one minute apart) from the source to the destination are lost. From these measurements we derive "failure epochs" on each path. The epoch begins when the third failed probe times out, and ends on the first successful reply from a subsequent probe. These epochs are the periods of time when the route between the source and destination may have failed.

This method of deriving failure epochs has a few limitations. Firstly, since we wait for three consecutive losses, we cannot detect failures that last less than 3 minutes. As a result, our analysis does not characterize the relative ability of overlays and route control to avoid such short failures. Secondly, ping packets may also be dropped due to congestion rather than path failure. Unfortunately, from our measurements we cannot easily determine if the losses are due to failures or due to congestion. Finally, the destination may not reply with ICMP echo reply messages within one second, causing us to record a loss. To mitigate this factor, we eliminate paths for which the fraction of lost probes is $> 10\%$ from our analysis. Due to the above reasons, the path failures we identify should be considered an over-estimate of the number of failures lasting three minutes or longer.

From the failure epochs on each end-to-end path, we compute the corresponding *availability*, defined as follows:

$$Availability \quad = \quad 100 \times \left(1 - \frac{\sum_i T_F(i)}{T}\right)$$

where, $T_F(i)$ is the length of failure epoch i along the path, and T is the length of the measurement interval (5 days). The total sum of the failure epochs can be considered the observed "downtime" of the path.

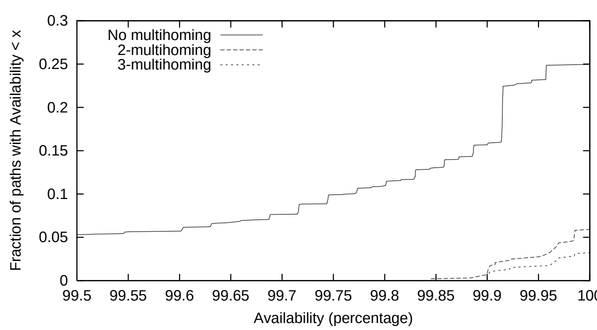


Figure 13: **End-to-end failures:** Distribution of the availability on the end-to-end paths, with and without multihoming. The ISPs in the 2- and 3-multihoming cases are the best 2 and 3 ISPs in each city based on RTT performance, respectively. k-Overlay routing, for any k, achieves 100% availability and is not shown on the graph.

In Figure 13, we show a CDF of the availability on the paths we measured, with and without multihoming. When no multihoming is employed, we see that all paths have at least 91% availability (not shown in the figure). Fewer than 5% of all paths have less than 99.5% availability. Route control with multihoming significantly improves the availability on the end-to-end paths, as shown by the 2- and 3-multihoming availability distributions. Here, for both 2- and 3-multihoming, we consider the combinations of ISPs providing the best round-trip time performance in a city. Even when route control uses only 2 ISPs, less than 1% of the paths originating from the cities we studied have an availability under 99.9%. The minimum availability across all the paths is 99.85%, which is much higher than without multihoming. Also, more than 94% of the paths from the various cities to the respective destinations do not experience any observable failures during the 5 day period (i.e.,availability of 100%). With three providers, the availability is improved, though slightly. Overlay routing may be able to circumvent even the few failures that route control could not avoid. However, as we show above, this would result in only a marginal improvement over route control which already offers very good availability.

6.2 Path Availability Analysis

Since the vast majority of paths did not fail even once during our relatively short measurement period, our second approach uses statistics derived from previous long-term measurements to ascertain availability. Feamster et al. collected failure data using active probes between nodes in the RON testbed approximately every 30 seconds for several months [9]. When three consecutive probes on a path were lost, a traceroute was triggered to identify where the failure appeared (i.e., the last router reachable by the traceroute) and how long they lasted. The routers in the traceroute data were also labeled with their corresponding AS number and also classified as border or internal routers. We use a subset of these measurements on paths between non-DSL nodes within the U.S. collected between June 26, 2002 and March 12, 2003 to infer failure rates in our testbed. Though this approach has some drawbacks (which we discuss later), it allows us to obtain a view of longer-term availability benefits of route control and overlay routing that is not otherwise possible from direct measurements on our testbed.

We first estimate the availabilities of different router classes (i.e., the fraction of time they are able to correctly forward packets). We classify routers in the RON traceroutes by their AS tier (using the method in [31]) and their role (border or internal router). Note that the inference of failure location is based on router location, but the actual failure could be at the *link or router* attached to the last responding router.

The availability estimate is computed as follows: If $\sum T_F^C$ is the total time failures attributed to routers of class C were observed, and N_d^C is the total number of routers of class C we observed on each path on day d,[6] then we estimate the availability of a router (or attached link) of class C as:

$$Availability_C \quad = \quad 100 \times \left(1 - \frac{\sum T_F^C}{\sum_d N_d^C \times one\_day}\right)$$

In other words, the fraction of time unavailable is the aggregate failure time attributed to a router of class C divided by the total time we expect to observe a router of class C in any path. Our estimates for various router classes are shown in Table 4.

AS Tier	Location	Availability (%)
1	internal	99.940
1	border	99.985
2	internal	99.995
2	border	99.977
3	internal	99.999
3	border	99.991
4	internal	99.946
4	border	99.994
5	internal	99.902
5	border	99.918

Table 4: **Availability across router classes:** Estimated availability for routers or links classified by AS tier and location. We consider a border router as one with at least one link to another AS.

To apply the availability statistics derived from the RON data set, we identified and classified the routers on paths between nodes in our testbed. We performed traceroute measurements approximately every 20 minutes between nodes

[6] The dataset only included a single successful traceroute per day. Therefore, we assumed that all active probes took the same route each day.

in our CDN testbed from December 4, 2003 to Dec 11, 2003. For our analysis we used the most often observed path between each pair of nodes; in almost all cases, this path was used more than 95% of the time. Using the router availabilities estimated from the RON data set, we estimate the availability of routes in our testbed when we use route control or overlay routing. When estimating the simultaneous failure probability of multiple paths, it is important to identify which routers are shared among the paths so that failures on those paths are accurately correlated. Because determining router aliases was difficult on some paths in our testbed,[7] we conservatively assumed that the routers at the end of paths toward the same destination were identical if they belonged to the same sequence of ASes. For example, if we had two router-level paths destined for a common node that map to the ASes `A A B B C C` and `D D D B C C`, respectively, we assume the last 3 routers are the same (since `B C C` is common). Even if in reality these routers are different, failures at these routers are still likely to be correlated. The same heuristic was used to identify identical routers on paths originating from the same source node. We assume other failures are independent.

A few aspects of this approach may introduce biases in our analysis. First, the routes on RON paths may not be representative of the routes in our testbed, though we tried to ensure similarity by using only using paths between relatively well-connected RON nodes in the U.S. In addition, we observed that the availabilities across router classes in the RON dataset did not vary substantially across different months, so we do not believe the difference in timeframes impacted our results. Second, there may be routers or links in the RON data set that fail frequently and bias the availability of a particular router type. However, since traceroutes are initiated only when a failure is detected, there is no way for us to accurately estimate the overall failure rates of all individual routers. Third, it is questionable whether we should assign failures to the last reachable router in a traceroute; it is possible that the *next* (unknown) or an even further router in the path is actually the one that failed. Nevertheless, our availabilities still estimate how often failures are observed at or just after a router of a given type.

Figure 14 compares the average availability using overlays and route control on paths originating from 6 cities to all destinations in our testbed. For overlay routing, we only calculate the availability of the paths for the first and last overlay hop (since these will be the same no matter which intermediate hops are used), and assume that there is always an available path between other intermediate hops. An ideal overlay has a practically unlimited number of path choices, and can avoid a large number of failures in the middle of the network.

As expected from our active measurements, the average availability along the paths in our testbed are relatively high, even for direct paths. 3-multihoming improves the average availability by 0.15-0.24% in all the cities (corresponding to about 13-21 more hours of availability each year). Here, the availability is primarily upper bounded by the availability of the routers or links immediately before the destination that are shared by all three paths as they converge.

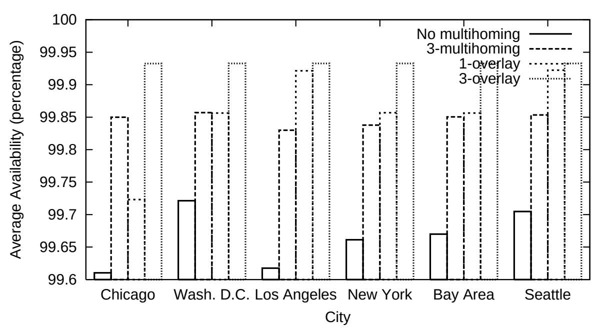


Figure 14: **Availability comparison:** Comparison of availability averaged across paths originating from six cities using a single provider, 3-multihoming, 1-overlays, and 3-overlays. ISPs are chosen based on their round-trip time performance.

In most cases, 1-overlays have slightly higher availability (at most about 0.07%). Since a 1-overlay has arbitrary flexibility in choosing intermediate hops, only about 2.7 routers are common (on average) between all possible overlay paths, compared to about 4.2 in the 3-multihoming case. However, note that a 1-overlay path using a single provider is more vulnerable to access link failures than when multihoming is employed. For example, the low availability

[7] We found that several ISPs block responses to UDP probe packets used by IP alias resolution tools such as Ally [29]

of the 1-overlay in Chicago is due to: (1) the chosen ISP (based on RTT performance) is a tier 4 network, which has internal routers with relatively lower availability, and (2) all paths exiting that provider have the first 5 hops in common and hence have a high chance of correlated failures. Finally, we see that using a 3-overlay usually makes routes only slightly more available than when using a 1-overlay (between 0.01% to 0.08%, excluding Chicago) . This is because at least one router is shared by all paths approaching a destination, so failures at that router impact all possible overlay paths. In summary, it is interesting to note that despite the greater flexibility of overlays, route control with 3-multihoming is still able to achieve an estimated availability within 0.08-0.10% (or about 7 to 9 hours each year) of 3-overlay.

7 Discussion

Next, we discuss observations made from our measurements and other fundamental tradeoffs between overlay routing and multihoming route control that are difficult to assess. We also comment on the limitations of our study.

Key observations. As expected, our results show that overlay routing does provide improved latency, throughput, and reliability over route control with multihoming. We found that overlay routing's performance gains arise primarily from the ability to find routes that are physically shorter (i.e. shorter propagation delay). In addition, its reliability advantages stem from having at its disposal a superset of the routes available to standard routing. The surprise in our results is that, while past studies of overlay routing have shown this advantage to be large, we found that careful use of a few additional routes via multihoming at the end-network was enough to significantly reduce the advantage of overlays. Since their performance is similar, the question remains whether overlays or multihoming is the better choice. To answer this, we must look at other factors such as cost and deployment issues.

Cost of operation. Unfortunately, it was difficult to consider the cost of implementing route control or overlays in our evaluation. In the case of multihoming, a stub network must pay for connectivity to a set of different ISPs. We note that different ISPs charge different amounts and therefore the solution we consider "best" may not be the most cost-effective choice. In the case of overlays, we envision that there will be overlay service offerings, similar to Akamai's SureRoute [1]. Users of overlays with multiple first hop choices (k-overlay routing in our analysis) must add the cost of subscribing to the overlay service to the base cost of ISP multihoming.[8] Using an overlay with a single provider (i.e., 1-overlays) would eliminate this additional cost, but our analysis shows that the performance gain is reduced significantly.

Deployment and operational overhead. Overlays and multihoming each have their unique set of deployment and performance challenges that our measurements do not highlight. Below, we consider the issues of ease of use and deployment, routing table expansion and routing policy violations.

Ease of use and employment. Overlay routing requires a third-party to deploy a potentially large overlay network infrastructure. Building overlays of sufficient size and distribution to achieve significantly improved round-trip and throughput performance is challenging in terms of infrastructure and bandwidth cost, as well as management complexity. On the other hand, since multihoming is a single end-point based solution, it is relatively easier to deploy and use from an end-network's perspective.

Routing table expansion due to multihoming. An important overhead of multihoming that we did not consider in this study is the resulting increase in the number of routing table entries in backbone routers. ISPs will likely charge multihomed customers appropriately for any increased overhead in the network core, thus making multihoming less desirable. However, this problem occurs only when the stub network announces the same address range to each of its providers. Since ISPs often limit how small advertised address blocks can be, this approach makes sense for large and medium sized stub networks, but is more difficult for smaller ones. Smaller networks could instead use techniques based on network address translation (NAT) to avoid issues with routing announcements and still make intelligent use of multiple upstream ISPs [13, 4].

Violation of policies by overlay paths. One of the concerns that overlay routing raises is its circumvention of routing policies instituted by intermediate ASes. For example, a commercial endpoint could route data across the relatively well-provisioned, academic Internet2 backbone by using an overlay hop at a nearby university. While each individual overlay hop would not violate any policies (i.e., the nearby university node is clearly allowed to transmit data across

[8] If the ISPs charge according to usage, then the cost of employing multiple ISP connections in the case of k-overlays may be higher or lower than the cost of using multiple connections in the case of k-multihoming.

Internet2), the end-to-end policy may be violated. While our analysis quantifies the number of routing policy violations, we did not consider their impact. Most Internet routing polices are related to commercial relationships between service providers. Therefore, it is reasonable to expect that the presence of an overlay node in an ISP network implies that the overlay provider and the ISP have some form of business agreement. This relationship should require that the overlay provider pay for additional expenses that the ISP incurs by providing transit to overlay traffic. Network providers would thus be compensated for most policy violations, limiting the negative impact of overlay routing.

Future changes to BGP. Thus far, we have discussed some important issues regarding overlays and route control in today's environment, but have not considered changes to BGP that may further improve standard Internet routing performance relative to overlays. For example, we only consider the impact of performance or availability-based route selection at the edge of the network. It is possible that transit ASes could perform similar route control in the future, thereby, exposing a superior set of AS paths to end networks. Another future direction is the development of new protocols for AS-level source-routing, such as NIRA [33], which allow stub networks greater control over their routes.

Limitations of the study. Our observations may be constrained by a few factors such as the size of our testbed, the coarse granularity of our performance samples, and our limited analysis of resilience. We discuss these issues in detail below.

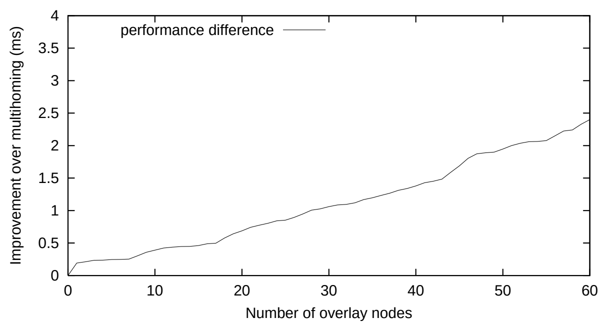


Figure 15: **Impact of overlay network size on round-trip performance:** This graph shows the mean difference between 3-overlays and 3-multihoming as overlay nodes are added.

Testbed size. In Figure 15 we compare the average RTT performance from 3-multihoming against 3-overlays, as a function of the number of intermediate overlay nodes available. The graph shows the RTT difference between the best 3-overlay path (direct or indirect) and best 3-multihoming path, averaged across all measurements as nodes are added one-by-one, randomly, to the overlay network. A different heuristic of adding nodes may yield different results. As the size of the overlay is increased, the performance of 3-overlays gets better relative to multihoming. Although the relative improvement is marginal, there is no discernible "knee" in the graph. Therefore it is possible that considering additional overlay nodes may alter the observations in our study in favor of overlay routing.

Granularity of performance samples. Our performance samples are collected at fairly coarse timescales (6 minutes intervals for round-trip time and 30 minutes for throughput). As a result, our results may not capture very fine-grained changes, if any, in the performance on the paths, and their effect on either overlay routing or multihoming route control. However, we believe that our results capture much of observable performance differences between the two path selection techniques for two key reasons: (1) our conclusions are based on data collected continuously over a week-long period, and across a fairly large set of paths, and (2) Zhang *et al.* observed that the "steadiness" of both round-trip time and throughput performance is at least on the order of minutes [34]. Other more recent measurements of round-trip times on similar paths as those in our testbed have shown mean intervals of several minutes between changes of 30% or more [4]. As such, we do not expect that a higher sampling frequency would yield significantly different.

Repair and failure detection. Our reliability analysis does not compare the relative ability of overlay routing and multihoming to avoid BGP convergence problems. For example, a peering link failure may affect routing between the peer ISPs until BGP re-converges. It is possible that some multihoming configurations cannot avoid such routing failures. We leave this comparison for future work.

8 Summary

Past studies have demonstrated the use of overlay routing to make better use of the underlying connectivity of the Internet than the current BGP-based system. However, BGP-based routing can benefit from the added capability of two important factors at end-networks: (1) additional access to end-to-end BGP routes via ISP multihoming, and (2) implementation of performance- and resilience-aware route control mechanisms to dynamically select among multiple BGP routes. In this paper, we have compared the relative benefits of overlay routing and intelligent route control and investigated possible reasons for the differences via an extensive measurement-based analysis. Our findings are as follows:

- Multihoming route control can offer performance similar to overlay routing. Specifically, overlays employed in conjunction with multihoming to 3 ISPs offer only about 3–12% better RTTs and 1–10% better throughput than route control in conjunction with multihoming to three ISPs. In fact, when overlays are constrained to a single first-hop ISP, they provide inferior performance relative to route control.
- The marginally better RTT performance of overlays comes primarily from their ability to select shorter end-to-end routes. Also, the performance gap between overlays and route control can be further reduced if, for example, ISPs implement mutually cooperative peering policies such as late-exit.
- While route control cannot offer the near perfect resilience of overlays, it can eliminate almost all observed failures on end-to-end paths. The path diversity offered by multihoming can improve fault tolerance of end-to-end paths by two orders of magnitude relative to the direct BGP path.

The results in our paper show that it is not necessary to circumvent BGP routing to achieve good end-to-end resilience and performance. These goals can be effectively realized by means of multihoming coupled with intelligent route control.

Acknowledgment

We would like to thank Olaf Maennel, Roberto De Prisco, Ramesh Sitaraman and Ravi Sundaram for their support and assistance with our experiments and data collection. We would also like to thank Nick Feamster and David Andersen for providing the RON failure data. Discussions and feedback from the following people have helped improve this work greatly: David Andersen, Hari Balakrishnan, Claudson Bornstein, Nick Feamster, Erich Nahum, Venkat Padmanabhan, Jennifer Rexford, Sambit Sahu and Hui Zhang. Finally, we thank our shepherd, Tom Anderson, and our anonymous reviewers for their valuable feedback and suggestions.

References

[1] Akamai Technologies. Akarouting (SureRoute). `http://www.akamai.com`, June 2001.

[2] Akamai Technologies. Edgescape. `http://www.akamai.com/en/html/services/edgescape.html`, 2004.

[3] A. Akella, B. Maggs, S. Seshan, A. Shaikh, and R. Sitaraman. A Measurement-Based Analysis of Multihoming. In *Proc. of ACM SIGCOMM '03*, Karlsruhe, Germany, August 2003.

[4] A. Akella, S. Seshan, and A. Shaikh. Multihoming Performance Benefits: An Experimental Evaluation of Practical Enterprise Strategies. In *Proc. of the USENIX 2004 Annual Technical Conference*, Boston, MA, June 2004.

[5] L. Amini, A. Shaikh, and H. Schulzrinne. Issues with Inferring Internet Topological Attributes. In *Proceedings of SPIE ITCOM*, August 2002.

[6] D. Andersen, H. Balakrishnan, M. Kaashoek, and R. Morris. Resilient Overlay Networks. In *Proc. of the 18th Symposium on Operating System Principles*, Banff, Canada, October 2001.

[7] N. Cardwell, S. Savage, and T. Anderson. Modeling TCP Latency. In *Proc. of IEEE INFOCOM 2000*, Tel Aviv, Israel, March 2000.

[8] F5 Networks. BIG-IP link controller. `http://www.f5.com/f5products/bigip/LinkController/`.

[9] N. Feamster, D. Andersen, H. Balakrishnan, and M. F. Kaashoek. Measuring the Effects of Internet Path Faults on Reactive Routing. In *Proc. of ACM SIGMETRICS 2003*, June 2003.

[10] N. Feamster, J. Borkenhagen, and J. Rexford. Guidelines for Interdomain Traffic Engineering. *ACM SIGCOMM Computer Communication Review*, October 2003.

[11] L. Gao. On Inferring Autonomous System Relationships in the Internet. *IEEE/ACM Transactions on Networking*, 9(6), December 2001.

[12] L. Gao and F. Wang. The Extent of AS Path Inflation by Routing Policies. In *Proc. of IEEE GLOBECOM 2002*, pages 2180–2184, 2002.

[13] F. Guo, J. Chen, W. Li, and T. Chiueh. Experiences in Building a Multihoming Load Balancing System. In *Proceedings of IEEE INFOCOM*, Hong Kong, March 2004. to appear.

[14] Y. Hyun, A. Broido, and k claffy. Traceroute and BGP AS Path Incongruities. Technical report, CAIDA, University of California, San Diego, 2003. `http://www.caida.org/outreach/papers/2003/ASP/`.

[15] IETF Traffic Engineering Working Group. `http://www.ietf.org/html.charters/tewg-charter.html`, 2000.

[16] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian. Delayed Internet routing convergence. *IEEE/ACM Transactions on Networking*, 9(3):293–306, June 2001.

[17] Z. Mao, R. Govindan, G. Varghese, and R. Katz. Route Flap Damping Exacerbates Internet Routing Convergence. In *Proc. of ACM SIGCOMM '03*, Karlsruhe, Germany, August 2003.

[18] Z. Mao, J. Rexford, J. Wang, and R. Katz. Towards an Accurate AS-Level Traceroute Tool. In *Proc. of ACM SIGCOMM '03*, Karlsruhe, Germany, August 2003.

[19] Nortel Networks. Alteon link optimizer. `http://www.nortelnetworks.com/products/01/alteon/optimizer/`.

[20] W. B. Norton. Internet Service Providers and Peering. In *Proceedings of NANOG 19*, Albuquerque, NM, June 2000.

[21] radware. Peer Director. `http://www.radware.com/content/products/pd/`.

[22] Y. Rekhter and T. Li. A Border Gateway Protocol 4 (BGP-4). Internet Request for Comments (RFC 1771), March 1995.

[23] M. Roughan, M. Thorup, and Y. Zhang. Traffic Engineering with Estimated Traffic Matrices. In *Internet Measurement Conference*, Miami, FL, November 2003.

[24] RouteScience Technologies, Inc. Routescience PathControl. `http://www.routescience.com/products`.

[25] S. Savage, A. Collins, E. Hoffman, J. Snell, and T. Anderson. The End-to-End Effects of Internet Path Selection. In *Proceedings of ACM SIGCOMM*, Boston, MA, September 1999.

[26] S. Savage et al. Detour: A Case for Informed Internet Routing and Transport. *IEEE Micro*, 19(1):50–59, 1999.

[27] A. Shaikh, J. Rexford, and K. G. Shin. Load-Sensitive Routing of Long-Lived IP Flows. In *Proc. of ACM SIGCOMM '99*, Cambridge, MA, September 1999.

[28] N. Spring, R. Mahajan, and T. Anderson. Quantifying the Causes of Internet Path Inflation. In *Proc. of ACM SIGCOMM '03*, August 2003.

[29] N. Spring, R. Mahajan, and D. Wetherall. Measuring ISP Topologies with Rocketfuel. In *Proc. of ACM SIGCOMM '02*, Pittsburgh, PA, August 2002.

[30] J. W. Stewart. *BGP4: Inter-Domain Routing in the Internet*. Addison-Wesley, 1999.

[31] L. Subramanian, S. Agarwal, J. Rexford, and R. H. Katz. Characterizing the Internet Hierarchy from Multiple Vantage Points. In *Proceedings of IEEE INFOCOM*, June 2002.

[32] H. Tangmunarunkit, R. Govindan, and S. Shenker. Internet Path Inflation Due to Policy Routing. In *SPIE ITCOM*, August 2001.

[33] X. Yang. NIRA: A New Internet Routing Architecture. In *Proc. of the ACM SIGCOMM Workshop on Future Directions in Network Architecture (FDNA)*, August 2003.

[34] Y. Zhang, N. Duffield, V. Paxson, and S. Shenker. On the Constancy of Internet Path Properties. In *Proc. of ACM SIGCOMM Internet Measurement Workshop (IMW)*, November 2001.