A Comparison of Overlay Routing and Multihoming Route Control

Aditya Akella Jeffrey Pang
Bruce Maggs Srinivasan Seshan
Carnegie Mellon University
{aditya, jeffpang,srini+, bmm}@cs.cmu.edu

Anees Shaikh
IBM T.J. Watson
Research Center
aashaikh@watson.ibm.com

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 BGP route control, when 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, throughput of 1MB TCP transfers, and path availability. We observe that the performance from route control employed in conjunction with multihoming to three ISPs (3-multihoming), is within 5-15% of that from 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.

Categories and Subject Descriptors

C.2 [**Computer Systems Organization**]: Computer-Communication Networks; C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design

General Terms

Measurement, Performance

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 [24, 25] 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 that contribute to the differences between overlay routing and BGP routing that have not been carefully evaluated in previous 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 [29]. 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 [24, 6]. 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[1].

In this paper, we compare overlays with end-point based deployments 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 com-

This work was supported by the Army Research Office under grant number DAAD19-02-1-0389. Additional support was provided by IBM.

SIGCOMM'04, August 30–September 3, 2004, Portland, OR.

[1]Several commercial vendors already enable such route selection (e.g., [19, 21, 23])

bination 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.

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 5–15% 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 2–10% 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 potentially be even further reduced 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. In Section 4, we analyze the RTT and throughput performance differences between route control and overlay routing and consider some possible reasons for the differences. In Section 5, we contrast the end-to-end availability offered by the two schemes. Section 6 discusses the implications of our results and presents some limitations of our approach. Finally, Section 7 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 [15, 16] and "inflation" of end-to-end paths due to BGP policies [27, 31]. Particularly relevant to our study are proposals for overlay systems to bypass BGP routing to improve performance and fault tolerance, such as Detour [24] and RON [6].

In the Detour work, Savage et al. [24] 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 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 [26], for example) and, intra- and inter-domain traffic engineering [18, 22, 14]. 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, 23]).

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").

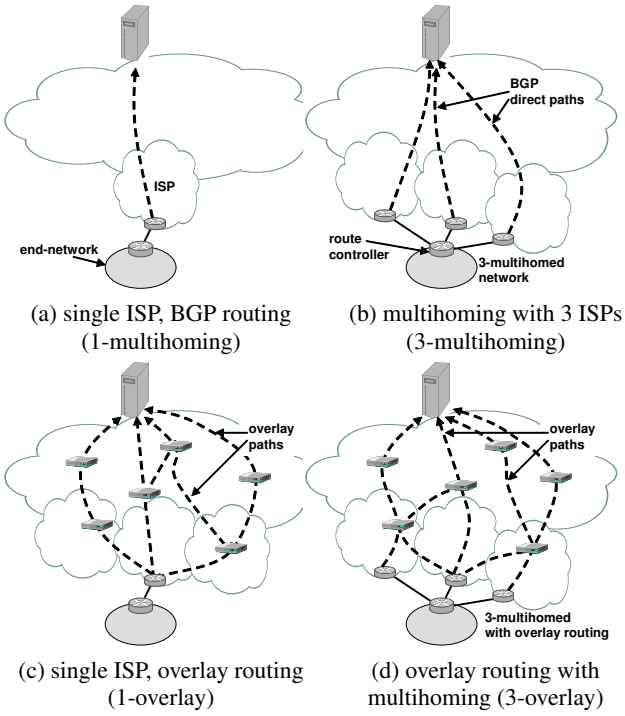
In a real implementation of multihoming route control, however, there are practical limitations on the ability of an end-network to track ISP performance, on the rate at which it can switch paths, and on the extent of control over incoming packets. However, recent work [4] shows that simple active and passive measurement-based schemes can be employed to obtain resilience and performance benefits 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 ($\sim$50 nodes)[2].

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-

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

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*.



(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 these cases, what is the extent of the performance difference?

2. What are the reasons for the performance difference? 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 failure rates that are comparable with overlay routing?

4. PERFORMANCE

In this section, we present our results on the relative performance benefits of multihoming route control compared with overlay routing. After describing our evaluation methodology in the next section, 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 [24] (Section 4.2). Then, we compare k-multihoming against 1-overlay routing, for $k \geq 1$ (Section 4.3). Here, we want to quantify the benefit of allowing greater flexibility in the choice of routes to end-systems via multihoming. Next, we contrast k-multihoming against k-overlay routing to understand the additional benefits offered by allowing end-systems almost arbitrary control on end-to-end paths, relative to multihoming (Section 4.4). Finally, we examine some of the underlying reasons for the performance differences (Sections 4.5 and 4.6).

4.1 Evaluation Methodology

We describe our approach, including descriptions of our testbed, how measurement data is collected, and our evaluation metrics.

4.1.1 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 a large 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 [24, 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 low likelihood of queuing compared to typical Internet paths. Although we do not claim that our measurements are completely representative, 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 [30]. The geographic distribution of the testbed nodes is illustrated in Figure 2(b). We consider emulated 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.

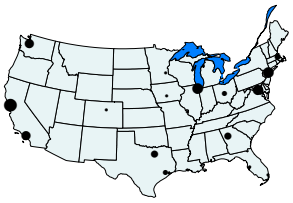
4.1.2 Data Collection

Our comparison of overlays and multihoming is based on observations drawn from two distinct data sets collected on our testbed. The first data set consists of active HTTP downloads of small objects (10KB) to measure the *turnaround times* between the pairs of nodes in our wide-area testbed. 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. The turnaround time samples are collected every 6 minutes between all node-pairs.

The second data set contains "throughput" measurements from active downloads of 1MB 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 (1MB) divided by the time between the receipt of the first and last bytes of the response data from the server (source). As we discuss below,

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.

this may not reflect the steady-state TCP throughput along the path.

Since our testbed nodes are part of a production commercial infrastructure, we limit the frequencies of our probes as described above. We ensure that each set of active probes between pairs of nodes are synchronized to occur within at most 30s of each other for the round-trip time data set, and at most 2 minutes for the throughput data set. For the latter, we ensure that an individual node is involved in atmost 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 US, the routes we probe, and consequently, our observations, are US-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.

4.1.3 Performance Metrics

Our performance comparison is based on two key metrics: round-trip time (RTT) and throughput. For each 6 minute interval, we build a weighted graph over all the 68 nodes where the edge weights are the RTT estimates obtained between the corresponding node-pairs. We use Floyd's algorithm to compute the shortest paths between all node-pairs. We do not prune the direct overlay edge in the graph before performing the shortest path computation. As a result, the shortest path 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 [24] 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 1MB object download frequency). The edge weights are the throughputs of the 1MB transfers (where throughput is obtained as defined above). Comparing the overlay throughput performance is complicated by the problem of estimating the end-to-end throughput for an 1MB 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, and the transfers are likely to remain in their slow-start phases, we use the small connection latency model developed in [7][3]. In the throughput data set, the paths we measured have a mean loss rate of 1% and median, 95th, 98th and 99th percentile loss rates of 0.004%, 0.01%, 0.7% and 20% respectively. Therefore a large fraction of the 1MB downloads hardly experience any packet losses.

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 an 1MB transfer. To combine loss rates on individual links, we follow the same approach as that described in [24]. We consider two possible combination functions. The first, called *optimistic*, uses the maximum observed loss on any individual overlay link, thus assuming 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 loss rates, assuming the losses on each link to be due to independent background traffic in the network[4].

Next, we make a few comments on the negligible loss rates we observe in our throughput measurements and their implications on our comparison between overlays and route control. First, since a majority of paths in the throughput dataset experience little loss, the difference between the two throughput estimates is likely to be negligible (as we show in the next section). Second, a simple calculation shows that, the lack of losses on the 1MB transfers suggest that the paths between our testbed nodes are provisioned well enough to accommodate upto 40Mbps bursts over very short time scales (< 100ms). Finally, the negligible loss implies that the throughput comparison between overlay and multihoming paths is likely to appear similar to the RTT comparison.

Also, due to the complexity of computing arbitrary length throughput-maximizing overlay paths, in our throughput comparison of overlay routing and routing control we only consider indirect paths comprised of at most two overlay hops.

4.2 1-Multihoming versus 1-Overlays

In this section, we compare the performance from overlay routing against using default routes via a single ISP (i.e., 1-overlay against 1-multihoming), along the same lines as [24].

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. Note that 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 70% compared to using direct BGP routes over a single ISP. Most cities see improvements of roughly 30%.

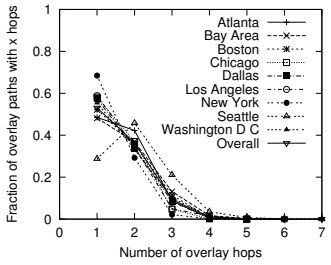
We show the distribution of overlay path lengths in Figure 3(b), where the direct (BGP) path corresponds to a single overlay hop. Notice that in most cities, the best overlay path is only one or two hops in more than 90% of the measurements. That is, the majority of the RTT performance gains in overlay networks are realized without requiring more than a single intermediate hop. Also, on an average 54% of the measurements, the best path from 1-overlays

[3]The initial congestion window size on our 1MB transfers was 2 segments. Also, for these transfers, there is no initial 200ms delayed ACK timeout on the first transfer.

[4]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	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	1.71
Wash D.C.	1.30

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



(b) 1-overlay path length

Figure 3: Round-trip time performance: Performance of 1-multihoming relative to 1-overlay routing, for RTT, 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. The best overlay path could be the direct path, in which case the number of overlay hops is 1.

City	Pessimistic estimate		Optimistic estimate	
	Throughput metric	*Fraction of indirect paths*	*Throughput metric*	*Fraction of indirect paths*
Atlanta	1.17	35%	1.17	35%
Bay Area	1.11	29%	1.11	29%
Boston	1.21	38%	1.21	38%
Chicago	1.15	32%	1.15	32%
Dallas	1.19	30%	1.19	30%
Los Angeles	1.13	16%	1.13	16%
New York	1.24	40%	1.24	40%
Seattle	1.35	44%	1.35	45%
Wash D.C.	1.14	34%	1.14	35%

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

coincides with the direct BGP path.

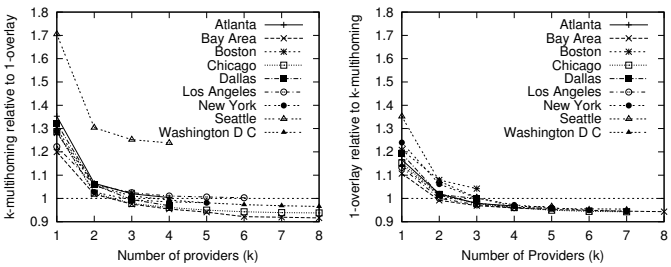
Throughput performance. In Table 1, we show the throughput performance of 1-overlays relative to 1-multihoming for both the pessimistic and the optimistic estimates. 1-overlays outperform 1-multihoming by about 11–35% in all cases, according to the pessimistic estimate. As we mention above, the results from the optimistic estimate are identical. 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, 1-overlay routing benefits from employing an indirect path in about 33% of the cases.

Summary. 1-Overlays offer significantly better round-trip time performance than 1-multihoming. However, the throughput benefits are less significant. Also, on a large fraction of the measurements, indirect 1-overlay paths offer better performance than direct 1-multihoming paths (46% for RTT and 33% for throughput).

4.3 k-Multihoming versus 1-Overlays

Compared to the previous analysis, in this section, we allow endpoints the flexibility of multihoming route control and compare the resulting performance against 1-overlays.

In Figure 4, 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



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

Figure 4: 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).

have a valid measurement over ISPs in the city.[5] We also note that in all but three cities, the best 3-multihoming providers according to RTT 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 4(a). Notice that 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 outperform overlays in many cities (with the exception of Seattle). Figure 4(b) shows relative benefits according to the pessimistic throughput estimate. Here, multihoming for $k \geq 3$ actually provides better throughput than 1-overlays. The results are similar for the optimistic computation and are omitted for brevity.

Summary. The performance benefits of 1-overlays are significantly reduced when the end-point is allowed greater flexibility in the choice of BGP paths via multihoming route control.

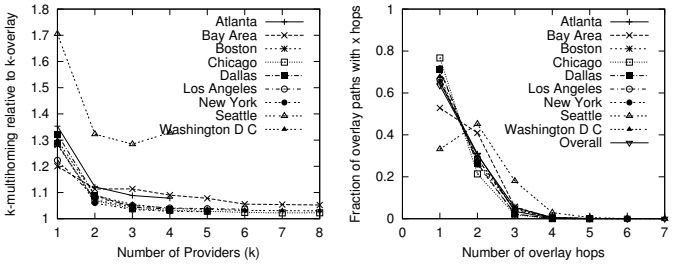
4.4 k-Multihoming versus k-Overlays

So far, we evaluated 1-overlay routing, where all overlay paths start by traversing a single ISP in the source city. Next, 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. Notice that the performance from the latter is at least as good as that from k-multihoming. The question we want to answer, then, is how much more advantageous overlays are if multihoming is already employed by the source.

Round-trip time performance. Figure 5(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 5–15% better RTT performance than the best multihoming solution in most of the cities in our study. In a few of cities the benefit is greater (e.g. Seattle and Bay Area). The performance gap between multihoming and overlays is even more marginal for $k \geq 4$.

Figure 5(b) shows the distribution of the number of overlay hops in the paths selected by 3-overlay routing, specifically. The best overlay path coincides with the best 3-multihoming BGP path in about 64% of the cases across all cities (Seattle and the Bay area are exceptions). Recall that the corresponding fraction for 1-overlay routing in Figure 3(b) was 54%. This shows that allowing overlay

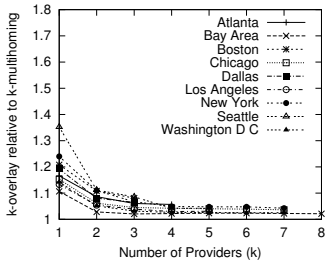
[5]Across all cities, an average of 10% of the time instants did not have a valid measurement across all providers in the city; nearly all of these cases were due to limitations in our data collection infrastructure, and not failed download attempts.



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

Figure 5: 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.

routing a greater choice of routes via more of the ISPs in a city actually causes it to select a *higher* fraction of direct BGP paths.



(a) Throughput improvement (pessimistic estimate)

City	Fraction of indirect paths
Atlanta	15%
Bay Area	11%
Boston	27%
Chicago	16%
Dallas	22%
Los Angeles	13%
New York	27%
Seattle	40%
Wash D.C.	20%

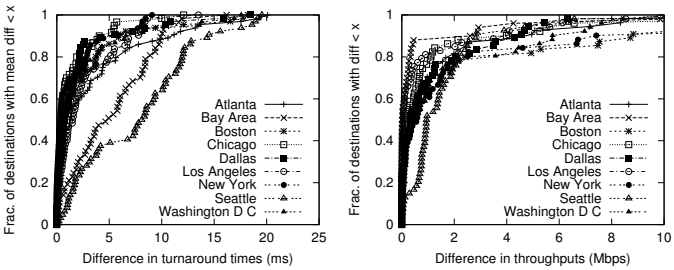
(b) Fraction of indirect paths in 3-overlay routing

Figure 6: 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 special case of k=3.

Throughput performance. Figure 6(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 2–12% of overlays, and the performance improves up to $k = 3$ or $k = 4$. In most cities, the throughput performance of 4-multihoming is within 4% of overlay routing. In Figure 6(b), we also show the fraction of measurements where an indirect 3-overlay path offers better performance than the direct 3-multihoming paths, for the pessimistic throughput estimate. On average, this fraction is about 23%. Notice that this is again lower than the corresponding percentage for 1-overlays from Table 1 ($\approx$ 33%). **Summary.** When employed in conjunction with multihoming, overlay routing offers marginal benefits over employing multihoming alone. In addition, k-overlay routing selects a larger fraction of direct BGP-based end-to-end paths, compared to 1-overlay routing.

4.5 Unrolling the Averages

In the previous sections, we presented averages of the performance differences for various forms of overlay routing and multihoming. 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 (2) by time-of-day or day-of-week effects.

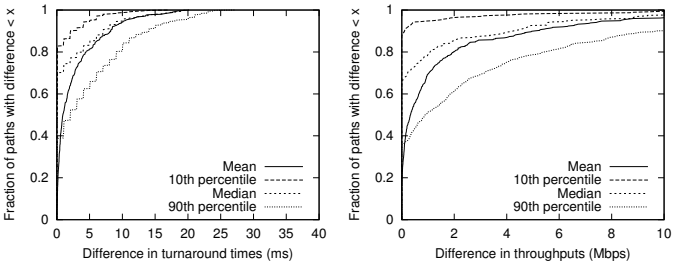


(a) Mean difference in round-trip times (b) Mean difference in throughputs (pessimistic)

Figure 7: 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 7(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 (except Seattle), the difference is greater than 15ms for less than 5% of the destinations.

In Figure 7(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 60–90% of the destinations. We also note that, for 2–20% of the destinations, the difference is in excess of 4 Mbps. Recall from Figure 6, however, that these differences result in an average relative performance advantage for overlays of less than 5–10% (for $k = 3$).



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

Figure 8: 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 8(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 overlay path across paths in all cities. In Figure 8(a) we see that the difference is fairly small, with more than 85% of the median RTT differences less than 10ms. The 90th percentile of the difference is marginally higher with roughly 10% greater than 15ms. The median throughput differences in Figure 8(b) are also relatively small – less than 500 kbps about 75%

of the time. Considering the upper range of the throughput difference, we see that a significant fraction (just under 40%) 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 consider 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 relatively worse during peak periods since overlay paths have greater freedom to avoid congested parts of the network. We did not see any discernible time-of-day effects in paths originating from a specific city, both in terms of RTT and throughput performances.

Similarly, we also examine weekly patterns to determine whether the differences are greater during particular days of the week, but again the differences were not significant for either RTT or throughput. We omit these result 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 BGP traffic engineering.

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

4.6 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 on the special case of $k = 3$. 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 inter-domain and peering policies are violated by indirect paths.

4.6.1 Propagation Delay and Congestion Improvements

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 (similar to [24]).

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.

In Figure 9, 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

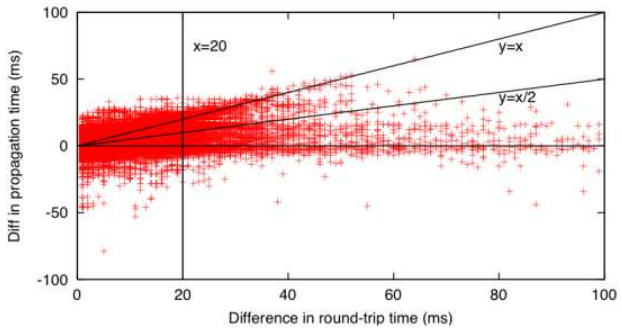


Figure 9: Propagation vs congestion: A scatter plot of the RTT improvement (x-axis) vs propagation time improvement (y-axis) of the indirect overlay paths over the direct paths.

improvements are due primarily to routing around congestion. We also found, however, that 72% 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 [24] 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 [24] could be due to the fact that Internet paths are better provisioned and less congested today than 3-4 years ago. However, they continue to be circuitous contributing to "inflation" in end-to-end paths [27].

Total fraction of lower delay overlay paths	36%	
	Fraction of lower delay paths	Fraction of all overlay paths
Indirect paths with $>$ 20ms improvement	4.7%	1.7%
Prop delay improvement $<$ x% of overall improvement (whenever overall improvement $>$ 20ms)		
$< 50\%$	2.2%	0.8%
$< 25\%$	1.7%	0.6%
$< 10\%$	1.2%	0.4%

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

Next, we focus further on measurements where indirect overlay paths offer significant improvement (> 20ms) over the best direct paths, to further investigate the relative contributions of propagation delay and congestion improvements. Visually, these are all point lying to the right of the $x = 20$ line in Figure 9. In Table 2 we present a classification all of the indirect overlay paths offering > 20ms RTT improvement. Recall that, in our measurement, 36% of the indirect 3-overlay paths had a lower RTT than the corresponding best direct path (Section 4.4, Figure 5 (b)). However, of these paths, only 4.7% 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 9. 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 ($\approx 0.8\%$).

Finally, if we consider the propagation delay of the best overlay

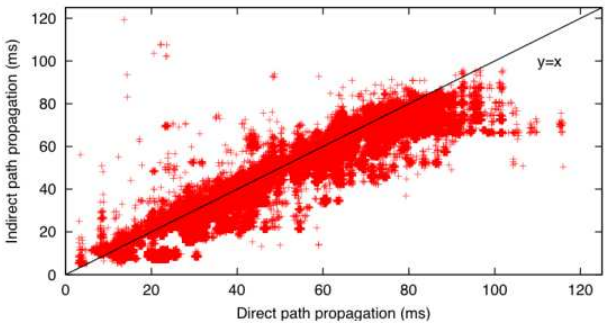


Figure 10: "Circuitousness" of routes: Figure plotting the propagation delay of the best indirect path (y-axis) against the best multihoming path (x-axis).

path versus the best multihoming path, we can get some idea of the relative ability to avoid overly "circuitous" paths. Figure 10 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[6].

4.6.2 *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 commonly represent business arrangements between ISPs [20, 27]. 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 [11] 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. [27] 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 deriving the AS path from traceroutes is that these router-level AS paths may be different than the actual BGP AS path [17, 5, 13], often due to the appearance of an extra AS number corresponding to an Internet exchange point or a sibling AS[7]. 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 [30] 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. E.g., 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 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 Improvement Paths**		
	%	*RTT Imprv (ms)*		%	*RTT Imprv (ms)*	
		Avg	90th		Avg	90th
Violates Inter-Domain Policy	**69.6**	8.6	17	**70.4**	37.6	46
Valley-Free Routing	64.1	8.5	17	61.6	36.7	45
Prefer Customer	13.9	9.1	17	15.3	51.4	76
Valid Inter-Domain Path	**22.0**	7.3	15	**19.4**	38.8	49
Same AS-Level Path	13.3	6.9	13	10.2	42.6	54
Earlier AS Exit	1.6	5.3	8	0.7	54.1	119
Similar AS Exits	6.1	6.4	12	5.8	39.3	53
Later AS Exit	5.6	7.8	14	3.8	45.6	57
Diff AS-Level Path	8.8	8.0	17	9.2	34.7	44
Longer than BGP Path	1.9	9.9	20	3.5	32.3	39
Same Len as BGP Path	6.4	7.6	16	5.5	36.2	45
Shorter than BGP Path	0.5	5.4	11	0.1	35.8	43
Unknown	**8.4**			**10.2**		

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 breaks down the indirect overlay paths by policy conformance. As expected, the majority of indirect paths (70%) violated either the valley-free routing or prefer customer policies. However, a large fraction (22%) appeared to be policy compliant. We subcategorize 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

[6]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.

[7]Two ASes identified as peers may actually be siblings [30, 10], 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.

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 indicates 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. Most indirect overlay paths violate common inter-domain routing policies. Also, we observed that a fair fraction of the policy-compliant overlay paths could be realized by BGP if ISPs employed cooperative peering policies such as late exit.

5. 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 downstream ASes [18]. 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 [15]. 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 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.

5.1 Active Measurements of Path Availability

Our first approach draws observations from 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, 2003 to January 28th, 2003. 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. These 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. Since we wait for three consecutive losses, we cannot detect failures that last less than 3 minutes despite recent work showing many failures lasting less than two minutes [9]. A significant fraction of these failures (more than two-thirds), however, occur in the networks of small ISPs (e.g., DSL providers). Since the nodes in our testbed are connected to large ISPs, we believe the likelihood of short outages that escape detection is low. Ping packets may also be dropped due to congestion rather than path failure. Unfortunately, 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:

$$Availability \quad = \quad 100 \times \left(1 - \frac{downtime}{probetime}\right)$$

where, $downtime$ is the sum total of the lengths of all failure epochs along the path and $probetime$ is the length of the measurement interval (5 days).

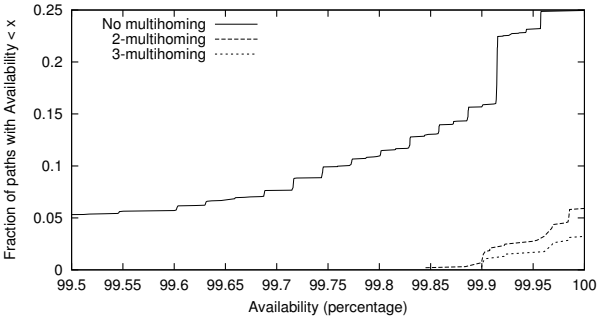


Figure 11: End-to-end failures: A CDF of the availability on the end-to-end paths, with and without multihoming. The ISPs in the 2- and 3-multihoming cases are the best 2 and 3 ISPs in each city based on RTT performance, respectively.

In Figure 11, 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 an availability less than 99.5%. 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 ISPs providing the best round-trip time performance in a city. Notice from this figure, that 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 further. 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.

5.2 Path Availability Analysis

Since the vast majority of paths did not fail even once during our measurement period, our second approach uses statistics derived from previous measurements taken over a considerably longer period [9]. Feamster et al. collected failure data using active probes between nodes in the RON testbed approximately every 30 seconds for several months. 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 the subset of these measurements on paths between non-DSL nodes within the U.S. collected between June 26, 2002 and March 12, 2003 to estimate failure rates in 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 [30]) and their role (border or internal router). The availability estimate is computed as follows: If $downtime_C$ is the total time failures attributed to routers of class C were observed, and $count_C^d$ is the total number of routers of class C we observed on each path on day d,[8] then we estimate the availability of a router of class C as:

$$Availability_C \quad = \quad 100 \times \left(1 - \frac{downtime_C}{\sum_d count_C^d \times one\_day}\right)$$

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 are shown in Table 4.

AS Tier	Location	Availability (%)
1	internal	99.939603
1	border	99.984791
2	internal	99.995445
2	border	99.976503
3	internal	99.998867
3	border	99.990682
4	internal	99.945846
4	border	99.994270
5	internal	99.901611
5	border	99.918372

Table 4: Availability across router classes: Estimated availability for routers classified by AS tier and location. A border router has at least one link to another AS.

Next, we performed traceroute measurements approximately every 20 minutes between nodes 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,[9] 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 set of routes on RON paths may not be representative of the set of routes in our testbed, though we believe they are somewhat similar given that we use only 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 time frames impacted our results. Second, there may be routers 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. In view of these observations, we believe that our availability estimates of our testbed paths may be somewhat approximate. However, we feel that our comparison of the relative ability of multihoming and overlay routing at improving failure tolerance is still reasonably accurate.

Figure 12 compares the average availability using the 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. This is not unreasonable, since 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 immediately before the destination that are shared by all three paths as they converge.

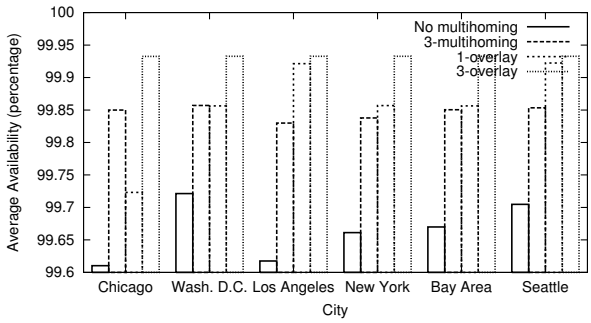


Figure 12: Availability comparison: Comparison of availability averaged across paths originating from six cities using a single provider, 3-multihoming, 1-overlays, and 3-overlays. ISPs are chosen based on their round-trip time performance.

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. In addition, the 1-overlay actually has a much lower average path availability in Chicago because: (1) the best performing provider 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. This suggests that the availability of routes in an overlay with a single initial provider is highly

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

[9]We found that several ISPs block responses to UDP probe packets used by IP alias resolution tools such as Ally [28]

dependent on how paths exit that provider. 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.

6. DISCUSSION

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

Key observations. As expected, our results show that overlay routing outperforms route control with multihoming for latency, throughput, and reliability. 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 come 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. While the number of ISPs needed probably dominates the cost of a multihoming setup, we should note that different ISPs charge different amounts and that 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. 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 large overlay network infrastructure. Building overlays of such magnitude for achieving improved round-trip and throughput performance is very challenging. 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. Also, most current overlays, especially RON [6] only facilitate communication between participants in the overlay. However, multihoming can be employed to communicate with any destination.

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 [12, 4].

Violation of policies by overlay paths. One of the concerns that overlay routing raises is its circumvention of routing policies instituted by intermediate AS hops. 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 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. So far, we have discussed some of the important issues regarding overlays and multihoming 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 performing performance or availability-based route selection from 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 [32], 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.

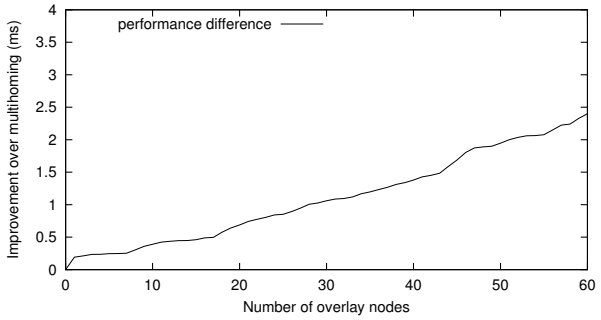


Figure 13: Impact of overlay network size on round-trip performance: This graph shows the mean difference between 3-overlays and 3-multihoming as overlay nodes are added.

Testbed size. In Figure 13 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 (direct or indirect) and best 3-multihoming path, averaged across all measurements as nodes are added randomly to the overlay network. 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 min intervals for round-trip time and 18 min 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 the round-trip times and throughput we measure is at least on the order of minutes [33]. As such, a higher sampling frequency is unlikely to yield very different results.

Repair and failure detection. Neither of our reliability analyses directly compares the ability of either scheme to avoid delayed convergence problems of BGP, where a link failure often affects routes within the vicinity of the failure. However, our static characterization using the failure probability of routers does reflect the amount of IP-level path diversity afforded by either scheme. Hence, we believe that this analysis does have indirect implications on the relative ability of either scheme to circumvent such convergence problems, since the availability of a diverse set of end-to-end paths should enable quick fail-over from such local problems.

7. 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) End-points can gain access to additional end-to-end BGP routes via ISP multihoming and (2) End-points can implement 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 any 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 conjuction with multihoming to 3 ISPs offer only about 5-15% better RTTs and 10% better throughput than route control in conjunction with multihoming to three ISPs. In fact, when overlays are constrained in their first hop connectivity, 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 are very grateful for the support and assistance of Roberto De Prisco and Ravi Sundaram of Akamai Technologies. 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.

8. 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] AKELLA, A., MAGGS, B., SESHAN, S., SHAIKH, A., AND SITARAMAN, R. A Measurement-Based Analysis of Multihoming. In *Proc. of ACM SIGCOMM '03* (Karlsruhe, Germany, August 2003).

[4] AKELLA, A., SESHAN, S., AND SHAIKH, A. Multihoming Performance Benefits: An Experimental Evaluation of Practical Enterprise Strategies. In *Proc. of the USENIX 2004 Annual Technical Conference* (Boston, MA, June 2004).

[5] AMINI, L., SHAIKH, A., AND SCHULZRINNE, H. Issues with Inferring Internet Topological Attributes. In *Proceedings of SPIE ITCOM* (August 2002).

[6] ANDERSEN, D., BALAKRISHNAN, H., KAASHOEK, M., AND MORRIS, R. Resilient Overlay Networks. In *Proc. of the 18th Symposium on Operating System Principles* (Banff, Canada, October 2001).

[7] CARDWELL, N., SAVAGE, S., AND ANDERSON, T. 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] FEAMSTER, N., ANDERSEN, D., BALAKRISHNAN, H., AND KAASHOEK, M. F. Measuring the Effects of Internet Path Faults on Reactive Routing. In *Proc. of ACM SIGMETRICS 2003* (June 2003).

[10] GAO, L. On inferring autonomous system relationships in the Internet. *IEEE/ACM Transactions on Networking 9*, 6 (December 2001).

[11] GAO, L., AND WANG, F. The Extent of AS Path Inflation by Routing Policies. In *Proc. of IEEE GLOBECOM 2002* (2002), pp. 2180–2184.

[12] GUO, F., CHEN, J., LI, W., AND CHIUEH, T. Experiences in Building a Multihoming Load Balancing System. In *Proceedings of IEEE INFOCOM* (Hong Kong, March 2004). to appear.

[13] HYUN, Y., BROIDO, A., AND K CLAFFY. Traceroute and BGP AS path incongruities. Tech. rep., CAIDA, University of California, San Diego, 2003. http://www.caida.org/outreach/papers/2003/ASP/.

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

2000.

[15] LABOVITZ, C., AHUJA, A., BOSE, A., AND JAHANIAN, F. Delayed Internet routing convergence. *IEEE/ACM Transactions on Networking 9*, 3 (June 2001), 293–306.

[16] MAO, Z., GOVINDAN, R., VARGHESE, G., AND KATZ, R. Route Flap Damping Exacerbates Internet Routing Convergence. In *Proc. of ACM SIGCOMM '03* (Karlsruhe, Germany, August 2003).

[17] MAO, Z., REXFORD, J., WANG, J., AND KATZ, R. Towards an Accurate AS-Level Traceroute Tool. In *Proc. of ACM SIGCOMM '03* (Karlsruhe, Germany, August 2003).

[18] N. FEAMSTER, J. BORKENHAGEN, AND J. REXFORD. Guidelines for Interdomain Traffic Engineering. *ACM SIGCOMM CCR* (October 2003).

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

[20] NORTON, W. B. 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] ROUGHAN, M., THORUP, M., AND ZHANG, Y. Traffic Engineering with Estimated Traffic Matrices. In *Internet Measurement Conference* (Miami, FL, November 2003).

[23] ROUTESCIENCE TECHNOLOGIES, INC. Routescience PathControl. `http://www.routescience.com/products`.

[24] SAVAGE, S., COLLINS, A., HOFFMAN, E., SNELL, J., AND ANDERSON, T. The End-to-End Effects of Internet Path Selection. In *Proceedings of ACM SIGCOMM* (Boston, MA, September 1999).

[25] SAVAGE, S., ET AL. Detour: A Case for Informed Internet Routing and Transport. *IEEE Micro 19*, 1 (1999), 50–59.

[26] SHAIKH, A., REXFORD, J., AND SHIN, K. G. Load-sensitive routing of long-lived IP flows. In *Proc. of ACM SIGCOMM '99* (Cambridge, MA, September 1999).

[27] SPRING, N., MAHAJAN, R., AND ANDERSON, T. Quantifying the Causes of Internet Path Inflation. In *Proc. of ACM SIGCOMM '03* (August 2003).

[28] SPRING, N., MAHAJAN, R., AND WETHERALL, D. Measuring ISP topologies with Rocketfuel. In *Proc. of ACM SIGCOMM '02* (Pittsburgh, PA, August 2002).

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

[30] SUBRAMANIAN, L., AGARWAL, S., REXFORD, J., AND KATZ, R. H. Characterizing the Internet Hierarchy from Multiple Vantage Points. In *Proceedings of IEEE INFOCOM* (June 2002).

[31] TANGMUNARUNKIT, H., GOVINDAN, R., AND SHENKER, S. Internet Path Inflation Due to Policy Routing. In *SPIE ITCOM* (August 2001).

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

[33] ZHANG, Y., DUFFIELD, N., PAXSON, V., AND SHENKER, S. On the Constancy of Internet Path Properties. In *Proc. of ACM SIGCOMM Internet Measurement Workshop (IMW)* (November 2001).