Topology Aware Overlay Networks

Junghee Han, David Watson, and Farnam Jahanian
Department of Electrical Engineering and Computer Science
University of Michigan
1301 Beal Avenue, Ann Arbor, Michigan, 48109-2122, USA
Email: {jungheeh, dwatson, farnam}@eecs.umich.edu

***Abstract*— Recently, overlay networks have emerged as a means to enhance end-to-end application performance and availability. Overlay networks attempt to leverage the inherent redundancy of the Internet's underlying routing infrastructure to detour packets along an alternate path when the given primary path becomes unavailable or suffers from congestion. However, the effectiveness of these overlay networks depends on the natural diversity of overlay paths between two endhosts in terms of physical links, routing infrastructure, administrative control, and geographical distribution. Several recent studies realized that a measurable number of path outages were unavoidable even with use of such overlay networks. This stems from the fact that overlay paths might overlap with each other when overlay nodes are selected without considering the underlying topology. An overlay network's ability to quickly recover from path outages and congestion is limited unless we ensure path independence at the IP layer. This paper proposes a novel framework for topology-aware overlay networks. In this framework, we expressly design overlay networks, aiming to maximize path independence without degrading performance. We develop measurement-based heuristics for 1) placement of overlay nodes inside an ISP and 2) selection of a set of ISPs. We base our analysis on extensive data collection from 232 points in 10 ISPs, and 100 PlanetLab nodes. On top of node placement, we present measurement-based verification to conclude that single-hop overlay routing performs as well as multi-hop routing with respect to both availability and performance. Our analysis results show that a single-hop overlay path provides the same degree of path diversity as the multi-hop overlay path for more than 90% of source and destination pairs. Finally, we validate the proposed framework using real Internet outages to show that our architecture is able to provide a significant amount of resilience to real-world failures.**

I. Introduction

A number of researchers have studied the stability, convergence, and end-to-end behavior of Internet routing protocols [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12]. These studies have revealed that the current underlying routing protocols are slow to react and recover from the failure of a link or router, and hence path failures and network congestion are visible to endhosts. This implies that although the Internet routing infrastructure is highly redundant, current underlying routing protocols fail to fully utilize alternative paths. Recently, overlay-based approaches have emerged as a means to circumvent these problems. An overlay network instantiates a virtual network on top of a physical network by deploying a set of overlay nodes above the existing IP routing infrastructure. Overlay nodes cooperate with each other to route packets on behalf of any pair of communicating nodes, forming an overlay network. Using these overlay networks, endhosts are able to select paths by themselves for better performance and availability without relying on the underlying IP routing infrastructure.

Existing overlay-based architectures, such as [13], [14], attempt to quickly recover from path failures and congestion problems by aggressively sending probes among overlay nodes at very short intervals. These networks trade the overhead of short-interval probes for prompt outage detection and recovery. In practice, however, several recent studies [13], [15] realized that approximately 40-50% of the path outages were still unavoidable even with the use of such overlay networks. This means that all alternate paths through overlay nodes suffered from path outages at the same time. This can happen because of loss and failure correlation between overlay paths at the underlying IP layer. There are many factors that contribute to the inter-dependency of path failure. For example, paths that travel across the same administrative domain can fail together due to a single configuration change or policy decision. Geographical adjacency can also be a factor. A failure at a Network Access Point (NAP) can affect all paths going through the NAP. Most of all, overlay paths that share the same physical links and/or routers are very likely to experience failure at the same time. Our prior study in [16] verified this argument by showing that a measurable amount of overlay paths overlap with each other when overlay nodes are randomly selected without considering the underlying topology. Hence, even with use of short interval probes, an overlay network's ability to quickly recover from path outages and congestion is limited unless we ensure that overlay paths go through disjoint IP layer paths.

In response to these observations, this paper proposes a novel framework for topology-aware overlay networks that enhances the availability and performance of end-to-end communication. This framework explicitly designs overlay networks to maximize path independence without degrading performance so that it can allow us to better utilize multi-homing at endpoints. To achieve this goal, we measure the diversity between different Internet Service Providers (ISPs) and also between different overlay nodes inside each ISP. Based on these measurements, we develop topology-aware node placement heuristics to ensure path diversity. This allows us to avoid path failures which are not avoidable using currently existing overlay-based approaches. In the measurement,

This work was supported in part by a research grant from the Defense Advanced Research Projects Agency, monitored by the U.S. Air Force Research Laboratory under Grant F30602-99-1-0527.

we rely on traceroute and ping probes collected from several vantage points in the Internet including *looking glasses* at ten major ISPs, and more than one hundred PlanetLab nodes [17]. In addition, we validate this framework based on real Internet failures. The primary contributions of this study are as follows:

- **A topology-aware overlay network framework to cope with path independence and improve availability and performance:** We explicitly design an overlay network to utilize path redundancy and maximize path independence between endhosts. The proposed topology-aware overlay framework is a novel approach to increasing the availability and performance of end-to-end communications. In the proposed framework, we deploy overlay nodes using off-line topology analysis rather than randomly deploying overlay nodes. Since operational topology change does not happen frequently,[1] this off-line node placement would only be updated over a long period as the Internet topology evolves. To accommodate transient topology changes due to congestion, link failures, or BGP instability, we provide flexibility in choosing overlay nodes on the fly, allowing the proposed framework to successfully detour faulty or congested paths.
- **Topology-aware node placement heuristics:** We propose several strategies to deploy overlay nodes while considering the underlying topology. With the proposed measurement-based guidelines, we can identify which and how many ISPs we need to deploy overlay nodes at. For instance, we observe that choosing three out of ten ISPs provides a similar degree of path diversity and latency benefit as deployment of all 10 ISPs in our experimental setup. In addition, we also present *clustering-based* heuristics to select a subset of overlay nodes inside the same ISP to maximize topological diversity between the nodes. Our evaluation shows that this node placement approach is able to recover from significantly more path outages than existing overlay networks.
- **A simple, but effective routing mechanism on top of the proposed topology-aware overlay architecture:** Our analysis results show that single-hop overlay paths provide the same degree of path diversity as multi-hop overlay paths for more than 90% of source and destination pairs. In addition, single-hop overlay paths improve latency for 90% of source/destination pairs compared with direct Internet paths. Therefore, we conclude that on top of topology-aware node deployment, single-hop overlay routing performs as well as multi-hop routing in terms of both availability and performance. In contrast to existing overlay networks [13], this single-hop overlay routing mechanism does not require a complicated routing protocol, whereas existing overlay solutions impose large overhead, and therefore are less scalable.
- **Evaluation of the proposed approach using real-world data:** We validate this proposed framework based on real Internet failures. In this evaluation, we show that the proposed approach is able to react and recover from about 87% of path outages while the existing overlay networks only recovered from about 50% of path outages. We construct our evaluation platform using 232 points from 10 ISPs and 100 PlanetLab nodes. The topological distribution of these collection points ensures that a broad range of ISPs are represented in our study. The evaluation platform captures real-world failure events and also logs if the overlay paths could avoid this failure. With this platform, we quantify how much network outage can be avoided by using the proposed framework.

Overall, our study provides guidance to administrators and researchers on how to incorporate topology considerations in designing overlay architectures.

The remainder of the paper is organized as follows. In the following section, we present background material and relate our work to prior stuides. Section III describes our measurement methodology and experimental results on node placement strategies. In Section IV, we show measurement-based verification of the effectiveness of single-hop routing. Section V presents an evaluation of our proposed framework. Finally, Section VI presents concluding remarks.

II. Background

The Internet infrastructure is inherently redundant. Specifically, the prevalence of redundant connections at the AS level has been discovered by many prior studies [18], [19]. Even fine-grained redundancy exists at the Internet link level. [20] quantified the topological redundancy of the Internet by showing the presence of a large number of disjoint paths at the link level. Although there is potential for utilizing this path redundancy, exploiting the redundancy requires a special framework such as overlay architectures.

Overlay networks attempt to leverage the inherent redundancy of the Internet's underlying routing infrastructure to detour packets along an alternate path when the given primary path becomes unavailable or suffers from congestion [13], [14], [21]. For instance, RON [13] nodes cooperate with each other to forward data on behalf of any pair of communicating nodes, forming an overlay network. If the underlying topology has physical path redundancy, it is often possible for RON to find alternative paths between nodes even if Internet routing protocols such as BGP do not use them. To find and use alternate paths, RON monitors the health of the underlying Internet paths between nodes, dynamically selecting paths that avoid faulty areas. Savage et al. [22] reported that in 30-80% of the cases, there is an alternate path with superior quality in terms of round-trip time, loss rate, and bandwidth.

However, the effectiveness of these overlay networks depends on the natural diversity of paths between two endhosts. Prior studies [23], [24], [16], [25] observed that two paths originating from different sources ISPs (or hosts) are very likely to experience a measurable amount of overlap. Although the Internet routing infrastructure is highly redundant, paths

[1]Peering relationship between ISPs are changing over a very long period—months or even longer.

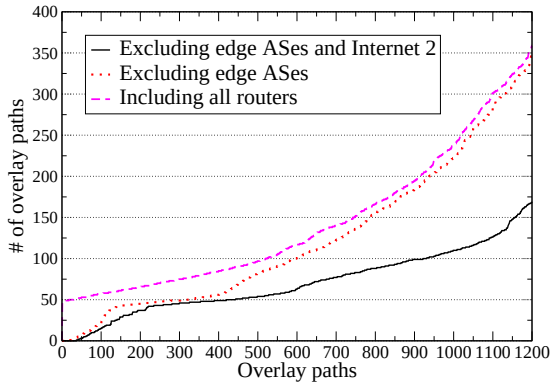


Fig. 1. Overlaps between overlay paths

actually taken by packets would not enjoy diversity. In addition, recent work [26] observed that overlay networks without multihoming might not be able to provide a high level of performance and reliability gains. This implies that current overlay networks with random deployment have limitations on ensuring path diversity.

To demonstrate this path diversity issue, we present one experimental result of our prior work [16] in Figure 1. For each pair of overlay paths, we examine whether or not these two overlay paths share any IP layer links. We repeat this procedure for all possible overlay path pairs. For each overlay path, l_i, we count how many other overlay paths share physical links with l_i. The x axis represents ith overlay path.[2] The top line in the graph corresponds to the case where two overlay paths share links/routers at the IP layer. The result shows that more than half of overlay paths share physical link/routers with more than 100 other overlay paths. This fact implies that even though two overlay paths are totally disjoint at the overlay layer, there is significant probability that they will overlap at the IP layer. The middle line in Figure 1, shows another analysis excluding the overlaps at the edge ASes. We also present another analysis excluding shared links/routers inside Internet2's Abilene network, represented as the bottom line. Overall, we observe that logically disjoint overlay paths between overlay nodes—placed in different ASes with distinct administrative control—are likely to share routers at the IP layer. This result explains why 40-50% of the path outages were still unavoidable even with use of overlay networks [15].

Inspired by these observations, this paper proposes a topology-aware overlay framework to maximize path independence for better availability and performance of end-to-end communication. To the best of our knowledge, there is no other work that develops a topology-aware overlay network by exploring the dependency between overlay paths. Several peer-to-peer studies [27], [28], [29] have attempted to construct topology-aware overlay networks for efficient data retrieval within the peer-to-peer domain. While such schemes can improve the efficiency of data retrieval, they are still not sufficient to guarantee fast response to path outages or increased performance. This is because they have only utilized informative hints about the nodes, not underlying IP topological information. For instance, each node in these peer-to-peer studies considers latency [27], [28] or lexicography [29] to choose neighbors, not the underlying IP path information. As a consequence, the selected neighbors for primary and backup paths might share a large amount of underlying links or routers whose failure can make primary and backup paths unavailable at the same time. In summary, applying these peer-to-peer techniques to our overlay network domain cannot guarantee resilience of end-to-end communication.

Recently, several studies including [30], [31], [32] have been proposed to develop a generic architecture that can be used for a variety of overlay applications with different requirements. In particular, [30] proposed an architectural element called a *routing underlay* that sits between overlay networks and underlying Internet. This architecture collects and tailors the Internet topology information, and answers application-specific queries, such as providing disjoint paths, based on the collected data. QRON [31] and X-bone [32] provide a general unified infrastructure which is shared by various overlay applications. Depending on an application-specific requirement, a different overlay topology is built on top of the proposed shared overlay infrastructure. However, these architectures do not explicitly consider an underlying Internet topology to construct their proposed architectures, which might limit the availability and performance gains. The proposed topology-aware node placement complements these studies by providing a guidance for strategic node deployment, and hence it can significantly enhance resilience of overlay network services.

III. Topology-Aware Node Placement

The effectiveness of overlay networks depends on the natural diversity of overlay paths. However, several recent studies, including [16], [25], [15], have observed a high possibility of overlaps among overlay paths when overlay nodes are randomly deployed. Hence, deploying overlay nodes, without consideration of the underlying IP topology, limits the network's ability to recover from path outages and network congestion. One solution to this problem might be to deploy overlay nodes on all routers at all ISPs and dynamically use the overlay nodes as backup paths depending on the source and destination. However, this is impractical due to the deployment overhead and associated economic costs.

In this paper, we propose several guidelines for *topology-aware node placement*. The goal is to select a subset of routers which are topologically diverse in order to provide independent paths for better availability and performance. We evaluate the diversity of all candidate ISPs and routers. By exploring these off-line statistics, a subset of ISPs and routers at which to place our overlay nodes are chosen. One possible exploration is to evaluate routers globally, ignoring ISP boundaries. This approach, however, might end up selecting

[2]We sort the index of an overlay path based on its y axis value. The y axis indicates the number of other overlay paths with whom l_i shares at least one router at the IP layer. For more detailed information, refer to [16].

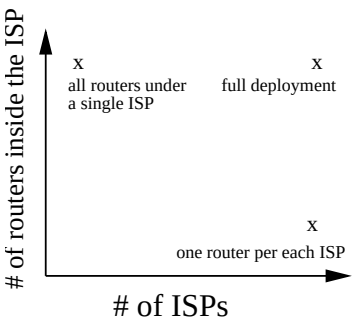
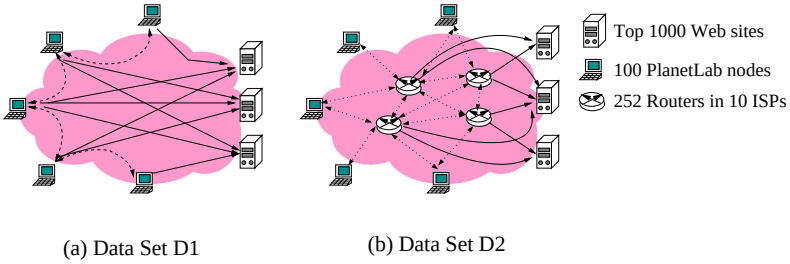


Fig. 2. Breadth vs. depth of node placement strategies



(a) Data Set D1 (b) Data Set D2

Fig. 3. Data sets

routers from only a single ISP. This is not a good design, because all selected routers are under the same administrative control and therefore internal problems of this administrative unit can affect all selected routers. Hence, this paper proposes hierarchical selection: we locally choose a subset of routers for each ISP first, and then we select a list of ISPs. In this approach, we can assure the diversity of the selected ISPs.

In the next two sections, we present several deployment strategies based on measured observations. In particular, we examine two related questions:

- Which ISPs and how many ISPs will we deploy the overlay nodes on? Would deploying at more ISPs provide significant gains?
- For each selected ISP, which and how many routers will we select to deploy overlay nodes at?

We can refer to the above two questions as breadth and depth of node placement strategies, as depicted in Figure 2. Full deployment (deploying overlay nodes at all possible places) is located at the right-top corner in this search space. Two extreme cases—1) selecting only one ISP but using all routers inside the selected ISP and 2) selecting all ISPs but using only a single router per ISP—are represented as the left-top and right-bottom corners, respectively. In the following sections, we perform measurement-based analysis to identify the most cost-effective depth and breadth of node placement search space. We conduct our analysis in a bottom-up manner; we first locally evaluate routers for each ISP (depth), and then compare different ISPs (breadth) using the results of the given depth analysis. Our proposed heuristics attempt to find a solution between these two extremes that approaches the availability and performance of full deployment but with lower deployment overhead.

A. Measurement Methodology

We define a *direct* path as the Internet path between endhosts without going through the overlay layer. On the other hand, the path via overlay nodes is referred to as an *indirect* path. To determine the quality of each overlay node, n_i, we use two metrics: path diversity and latency. For path diversity, we compute the number of shared routers between the direct path and the indirect path through n_i. With the latency metric, the quality of n_i is defined as the round-trip time difference between the direct path and the indirect path via n_i. To gather the direct and indirect path information, we rely on two data sets, D_1 and D_2, respectively, as described below.

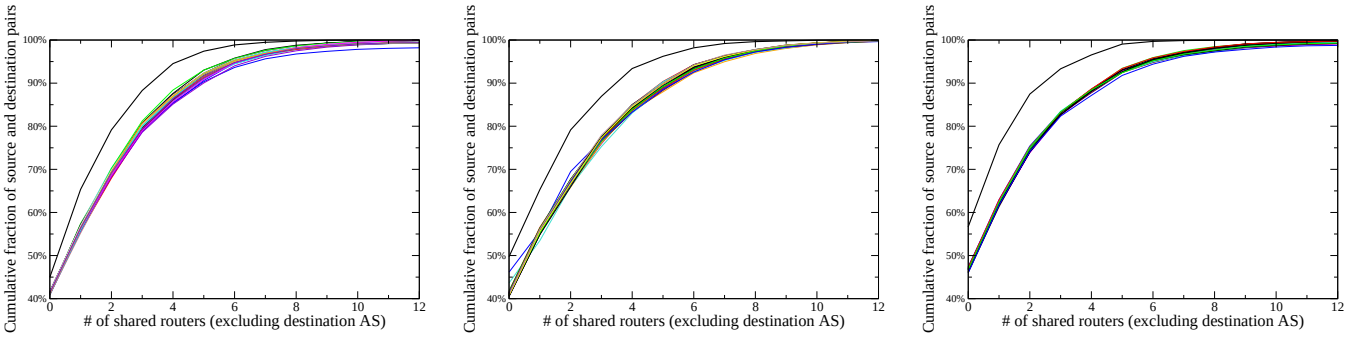
a) Data set $\boldsymbol{D_1}$: To measure the direct Internet paths, we collect traceroute and ping data from 100 PlanetLab nodes at stub networks. PlanetLab is an open, globally distributed testbed for deploying and accessing planetary-scale network services [17]. We consider these nodes as our target customer networks/endhosts and run traceroute from these points to 1) every other PlanetLab node and 2) top 100 Web sites, as shown in Figure 3(a).

b) Data set $\boldsymbol{D_2}$: To evaluate the impact of the choice of overlay nodes, we collect another set of traceroute and ping data from topologically and geographically diverse vantage points located in various ISPs. We take advantage of looking glasses offered by 10 ISPs: 6 different tier-1 ISPs—Cable&Wireless, Sprint, Qwest, Level 3 Communications, Teleglobe, and Global Crossing—and 4 small size ISPs. Looking glasses are publicly accessible Web sites provided by ISPs, where customers can measure performance and availability statistics using several utilities such as traceroute, ping, and BGP data. For example, each looking glass provides a tool for triggering traceroutes from several different routers inside the ISP to arbitrary destinations. Using the looking glasses, we can access 232 routers from 10 ISPs. We consider these routers as possible places to deploy overlay nodes. Note that routers within the ISP are also geographically distributed. We trigger traceroutes from these points to 1) 100 PlanetLab nodes and 2) top 100 Web sites, as shown in Figure 3(b).

B. Placement of overlay nodes inside an ISP network

In this section, we attempt to answer the question: which and how many routers should we select from each ISP? We evaluate overlay nodes inside a single ISP with respect to both path diversity and latency.

To measure path diversity, we rely on traceroute data included in the two data sets, D_1 and D_2. We count the number of overlapping routers between the direct path and the indirect path through a given overlay node as a path diversity metric. We apply this procedure to the 100*(100+100) pairs of source and destination hosts. In Figure 4, we show the measured path diversity for only 3 ISPs due to space limitations. The



(a) Cable&Wireless (b) Level 3 Communications (c) Qwest

Fig. 4. Comparing path diversity of different overlay nodes within the same ISP

ISP		# of clusters (α = correlation threshold)		
Name	# of router	α = 0.95	α = 0.9	α = 0.5
C&W	30	11	4	3
Global Crossing	35	7	1	1
Level 3	24	10	2	1
Qwest	22	16	13	6
Sprint	28	16	4	3
Teleglobe	28	24	24	3

TABLE I
STATISTICS OF CLUSTERING: PATH DIVERSITY

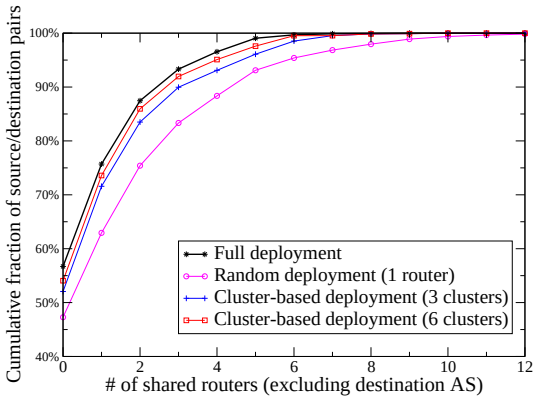


Fig. 5. Cluster-based deployment: Qwest

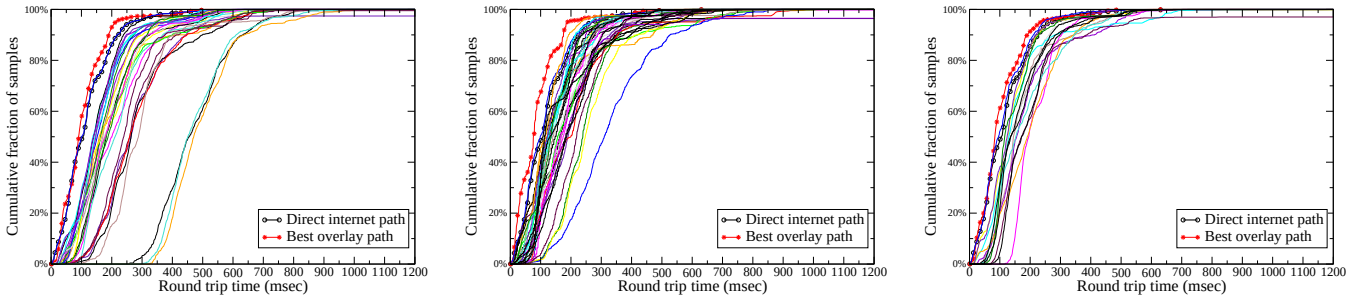
x axis indicates the number of shared routers and the y axis represents the cumulative fraction of source and destination pairs. Each line, except the leftmost line, represents the case where the corresponding overlay node is statically selected for all samples. On the other hand, the leftmost line represents the optimal case where we intelligently (or dynamically) select the optimal overlay node depending on the source and destination. From this experiment, we found that:

- Each line in Figure 4 is very close to every other line. This indicates that regardless of which router we select, the overall path diversity provided by the individual overlay node is almost the same. Any single overlay node does not provide the best path diversity for every pair of sources and destinations.
- On the other hand, the dynamic selection of overlay nodes, represented as the leftmost line in each graph of Figure 4, shows much better path diversity than the static cases. This implies that although overlay nodes are located within the same ISP, the paths taken from these nodes are different. Hence, it is important to have more than one overlay node within the same ISP. This provides flexibility for choosing a proper overlay node depending on an individual source and destination pair.

The above observations lead to the following questions: 1) how many routers are needed to provide optimal or nearly optimal diversity? and 2) which routers within the same ISP should we choose? To answer these questions, we present a *clustering-based* heuristic as described below. With this heuristic, we identify a subset of overlay nodes which provide a similar degree of path diversity as the optimal case.

In this heuristic, we first examine path diversity patterns of overlay nodes within the same ISP and categorize them into different clusters based on their patterns. For instance, two overlay nodes fall into the same cluster when their path diversity patterns over 100*(100+100) source and destination pairs are similar to each other. As a metric to determine the similarity between two overlay nodes, we use the correlation of path diversity over all source and destination pairs, as formulated below.

- S: a set of source hosts (i.e., 100 hosts)
- D: a set of destination hosts (i.e., 200 hosts)
- N: the number of source and destinations pairs (i.e., 100*200)
- $I_{s,d}$: the number of overlapping routers between the indirect overlay path through overlay node i and direct path from the source, s, to the destination, d
- E(I), STD(I): expectation and standard deviation values



(a) Cable&Wireless (b) Level 3 Communications (c) Qwest

Fig. 6. Comparing latency of different overlay nodes within the same ISP

of the random variable I, respectively

$$\begin{aligned} \text{Corr}(I,J) &= \frac{\text{E}(IJ)-\text{E}(I)\text{E}(J)}{\text{STD}(I)\text{STD}(J)} \\ \text{E}(I) &= \frac{1}{N}\sum_{s\in S,\ d\in D} I_{s,d} \\ \text{STD}(I)^2 &= \frac{1}{N}\sum_{s\in S,\ d\in D} (I_{s,d}-\text{E}(I))^2 \\ \text{E}(IJ) &= \frac{1}{N}\sum_{s\in S,\ d\in D} I_{s,d}J_{s,d} \end{aligned}$$

If the correlation of two overlay nodes is higher than a threshold value, α, we consider these two overlay nodes to provide similar patterns of path diversity. In Table I, we show the statistics of clustering for 6 tier-1 ISPs with different values of the correlation threshold. For example, the 30 routers inside Cable&Wireless are clustered into 4 groups with α = 0.9. Note that we do not claim that some absolute value of α is good or bad. We are more concerned about the number of clusters, which is directly related to economic cost limitations. In this paper, we use various α values to control the number of clusters.

Based on these clustering statistics, we propose a method to choose overlay nodes inside ISPs: randomly selecting one overlay node out of each cluster might perform as well as having all overlay nodes deployed. To evaluate this heuristic, we compare the path diversity of our proposed cluster-based deployment with the ideal full deployment and random deployment as references. Without loss of generality, we present the result from one ISP (i.e., Qwest) in Figure 5. In this graph, we show full, 3-cluster (α=0.5), 6-cluster (α=0.92), and random deployments. As shown in this figure, we observe that cluster-based heuristics provide significant path diversity gains compared to random deployment, and furthermore this clustering-based method is able to perform as well as full deployment.

While the above heuristics allow us to ensure path diversity, we still want to be able to limit the latency overhead of our overlay nodes. Hence, we conduct further analysis to evaluate the performance of each overlay node. We rely on ping probes included in the two data set, D_1 and D_2 to obtain roundtrip times between two nodes. For each source and destination pair, the roundtrip time of the overlay path through an overlay node n_i is defined as the sum of the two roundtrip times: between the source and n_i, and between n_i and the destination. In Figure 6, we show the cumulative distribution of roundtrip times for each overlay node. Our major findings are:

- In contrast to path diversity, overlay nodes inside the same ISP show very different patterns of latency. European or Asian overlay nodes present much longer latency for most destinations than the ones in the U.S.A. For example, the three rightmost lines in Figure 6(a) show the overall latency distribution of overlay nodes in Europe and Asia. Hence, selection of overlay nodes becomes more critical with respect to latency rather than path diversity.
- The dynamically selected overlay path provides even better latency than direct Internet paths, which is consistent with observations from prior work [13], [14].

To identify a subset of routers which provide the best overall latency, we apply the same clustering method as explained above with the exception that we use round-trip time instead of path diversity to calculate the similarity between overlay nodes. Table II shows the statistics of latency-based clustering for 6 ISPs with different correlation threshold values. From each cluster, we randomly select one router as an overlay node. By applying such cluster-based heuristics, we can achieve a similar degree of performance as the optimal case or even provide better performance than direct Internet paths. Figure 7 shows an example of our analysis results for Cable&Wireless. In this figure, we observe that the performance of the proposed clustering-based deployment is very close to full deployment and the direct Internet path. More interestingly, we see that clustering deployment always provides better performance than any static selections, represented as dotted lines in this figure.

Until now, we have examined path diversity and latency separately. However, satisfying both latency and path diversity goals at the same time can be very difficult, just like any optimization problem. For example, one router might provide good path diversity but significantly degrade performance. Hence,

ISP		# of clusters (α = correlation threshold)		
Name	# of router	α = 0.9	α = 0.7	α = 0.5
C&W	30	3	2	1
Global Crossing	35	7	2	1
Level 3	24	15	2	1
Qwest	22	18	7	4
Sprint	28	27	3	2
Teleglobe	28	24	4	1

TABLE II
STATISTICS OF CLUSTERING: LATENCY

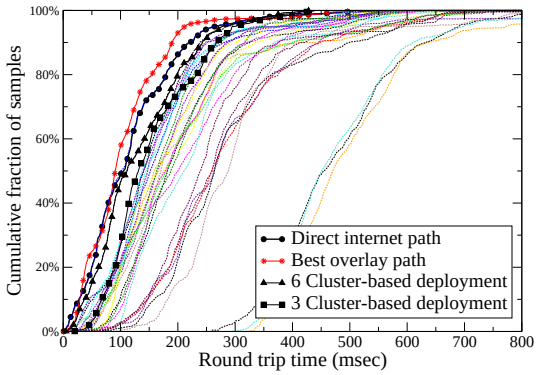


Fig. 7. Cluster-based deployment: C&W

we propose an algorithm for node selection combining latency- and path-based clustering methods, as described in Figure 8. With this combined method, we attempt to provide near optimal path diversity and latency gains. In this algorithm, we assign a tuple (c_p, c_l) to each router, where c_p and c_l indicate the cluster number with respect to path diversity and latency, respectively. Then, we perform the iterative search to select one router from each cluster. We use latency-based clusters first. We select one router from each latency cluster (line 1-4). Next, we check if diversity clusters are covered by the selected routers. If not, we add a randomly selected router from the uncovered diversity-based clusters. We can optimize this algorithm to reduce the total number of selected routers by replacing line 2 with a new line: "`choose one router whose diversity-cluster is not yet covered`" instead of selecting randomly.

C. Choosing a set of ISP networks

In the previous section, we examined the problem of placing a set of overlay nodes within a single ISP. Now we present the method for identifying which ISPs and how many ISPs to deploy overlay nodes at. With this methodology, we are able to choose the proper subset of ISPs which provide a similar degree of path diversity and latency benefit as full deployment. This study provides administrators with systematic tools to maximize topological diversity between ISPs within their budget and ISP contract limitations. We demonstrate the methodology through a representative data set; relying on the

1. **FOR** cluster $c_l \in$ *Set-Of-Clusters-Latency*
2. randomly choose one router from cluster c_l
3. insert the selected router into *Set-Of-Routers*
4. **END**
5. **FOR** cluster $c_p \in$ *Set-Of-Clusters-Diversity*
6. **IF** at least one router in *Set-Of-Routers* is in c_p
7. goto next;
8. **ELSE**
9. randomly choose one router from cluster c_p
10. insert the selected router into *Set-Of-Routers*
11. **END**

Fig. 8. Combining path diversity and latency

two data sets, D_1 and D_2, we evaluate 10 different ISPs with respect to path diversity and latency.

It is not straightforward to directly compare ISPs. The path diversity and latency of each ISP is not simply calculated from one router; rather we need to compute an abstraction of each ISP from an analysis of several different routers inside each ISP. To address this issue, we choose k routers per ISP ($1 \leq k \leq 10$) as a representative of each ISP by using the cluster-based methods described in the previous section. This k-$cluster$ driven abstraction is consistent with the proposed node placement strategy described before—we select the ISP and locally choose the k most diverse routers for that selected ISP.

First, we examine the path diversity of individual ISPs. We conduct analysis varying k from 1 to 10. Without loss of generality, we show the analysis results when $k = 3$ in Figure 9. Each line except the leftmost line represents the case where we statically choose only the corresponding ISP[3] for backup paths for all source/destination pairs. The leftmost line represents the optimal case where we intelligently (or dynamically) select one among 10 ISPs depending on the source and destination. Each line in Figure 9 is very close to every other line. This indicates that the overall path diversity provided by the individual ISP is almost the same, so any single ISP does not provide the best path diversity for every source and destination pair. On the other hand, the dynamic selection of ISPs, represented as the leftmost line, shows much better path diversity than the static cases. Hence, we believe that it is important to carefully choose more than one ISP in order to provide better path diversity.

Now, we examine in more detail the issue of determining which and how many ISPs we need to select. To answer this question, we study the path diversity gains by increasing the number of selected ISPs from 1 to 10. For 1-ISP selection, we use the best ISP as a representative. For 2-ISP selection, we evaluate the benefit of each pair of ISPs for all $\binom{10}{2}$ options. Among the $\binom{10}{2}$ options, we choose the best pair of ISPs as a representative. We repeat this procedure from 1-ISP selection to 10-ISP selection. Figure 10(a) shows the benefit of n-ISP selection. The x and y axes represent the number of ISPs (i.e.,

[3]Inside the selected ISP, we dynamically choose the best router out of k routers depending on the source and destination.

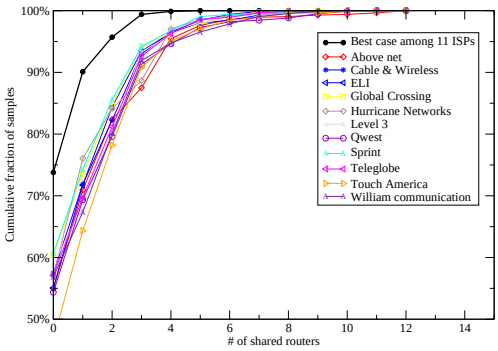


Fig. 9. Comparison of different ISPs

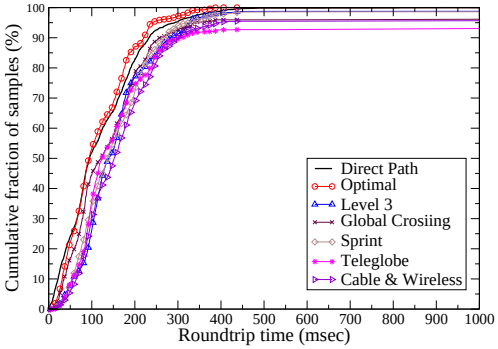


Fig. 11. Would choosing different ISPs provide latency gains?

$1 \leq n \leq 10$) and the number of routers per ISP (i.e., $1 \leq k \leq 6$). The z axis indicates the fraction of source and destination pairs which do not experience any overlapping between the direct Internet path and indirect overlay paths through the selected ISPs and routers. We observe that including more ISPs provides better path diversity, which is not surprising. Changing from 1-ISP to 2-ISP gives the most significant gain in our experiment. In our data set, choosing Sprint and Teleglobe gives the most cost-effective benefits. However, having more than three ISPs provides only marginal gains in our study. We also found that choosing four routers per each ISP is most cost-effective regardless of the number of selected ISPs. Overall, this experimental result implies that a subset of ISPs might be able to provide a similar degree of path diversity as full deployment, and for each selected ISP, we need to deploy overlay nodes at only a subset of routers.

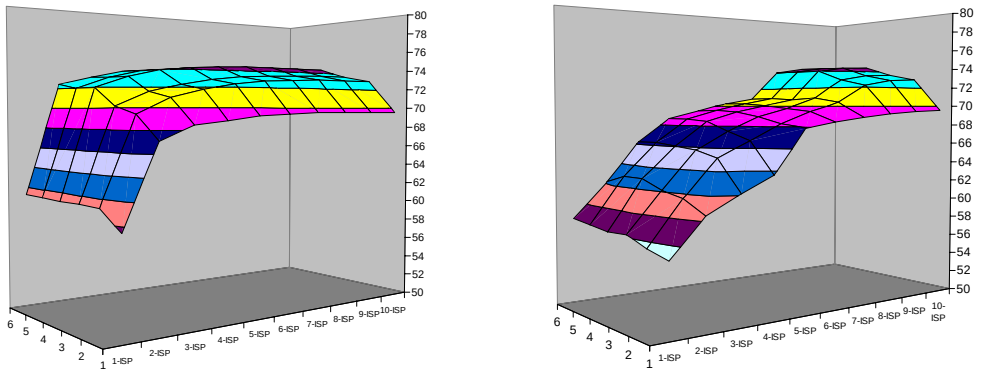
Using the above analysis, we are able to determine a set of ISPs to deploy overlay nodes at. However, one question might arise: how important is it to get the best set of ISPs? Does the random choice of ISPs degrade path diversity in a measurable amount? To answer this question, we repeat the same analysis as above except that we use the median set of ISPs instead of the best one from each n-ISP selection. Figure 10(b) shows the results from median selection. For 2-, 3-, or 4-ISP selections, choosing the median set instead of the best one measurably degrades path diversity. This implies that careful selection of ISPs is important, especially when only a small number of ISPs are selected; there typically exist "bad" ISPs and we need to filter them out. In contrast, with more than 4-ISP selection, the choice of ISPs does not matter much. This is due to the fact that as we increase the number of ISPs, it is more likely that the "good" ISP will be included in the median combinations. Overall, this analysis provides administrators with systematic tools to choose the proper set of ISPs within their budget and ISP contract limitations.

Finally, we examine how the selection of an ISP affects latency. We select 3 routers from each ISP as a representative using our latency-based clustering method, and compare an individual ISP for 100*(100+100) source and destination pairs. In Figure 11, each line except the two leftmost lines indicates selection of an individual ISP. The leftmost line represents the optimal case where we intelligently (or dynamically) select the best ISP depending on the source and destination. The second leftmost line corresponds to direct Internet paths. In contrast to path diversity, 1-ISP selection can perform as well as the full deployment and the direct Internet paths. Hence, we believe that the choice of ISP does not matter with respect to latency. This observation is the exact opposite of the previous results in Section III-B. Recall that in that section, we performed intra-ISP analysis—we evaluated individual routers inside the same ISP. In that analysis, we found that each router shows very different patterns of latency and hence selection of overlay nodes inside the ISP becomes critical. This is mainly due to the broad range of their geographical locations. Overall, we conclude that with respect to latency, which ISP each router belongs to is not critical. Instead, other factors such as geographical location are more important.

IV. Single vs. Multi-hop Overlay Routing

Existing overlay networks such as [13], [14] require additional complicated routing mechanisms at the overlay layer on top of IP routing mechanisms. For example, RON[13] adopts a link-state routing protocol between overlay nodes with short-interval probes, which hampers scalability. This architecture trades the overhead of such short-interval probes and additional routing for prompt outage detection and recovery. However, the authors of RON observed that a majority of the path outages in their experimental results were avoided by detouring through only a single overlay node. In this study, we complement their studies by providing measurement-based verification instead of anecdotal observation to conclude that single-hop overlay routing performs as well as multi-hop routing in terms of both availability and performance. By adopting single-hop routing, we need not exchange routing-related data between overlay nodes. In addition, there is no extra delay from packets transiting multiple overlay nodes, making this solution more scalable.

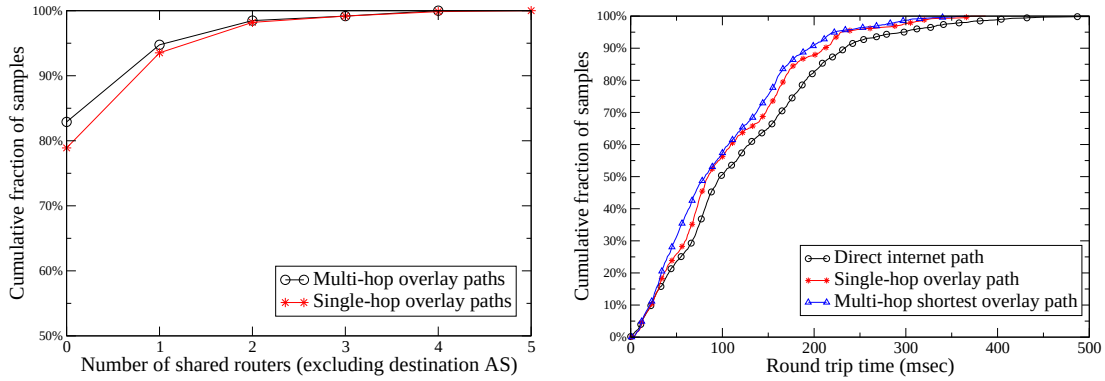
The first part of our analysis compares the path diversity provided by a single-hop overlay path with the diversity provided by multi-hop overlay paths. In this analysis, we again rely on the two data sets, D_1 and D_2, described in Section III-A. For each pair of the source and destination, s and d, direct path $DP_{s,d}$ corresponds to the traceroute from s to d. The single-hop overlay path, $SP^i_{s,d}$, through overlay



(a) 100th Percentile (the best selection)

(b) 50th Percentile (median selection)

Fig. 10. Would having more ISPs provide significant path diversity gains?



(a) Path diversity

(b) Latency

Fig. 12. Single vs. multi-hop overlay paths

node i is composed of two measurements: one from s to i and the other from i to d. For the multi-hop path, we compute the *optimal* overlay path, $MP_{s,d}$, by assuming the best-case scenario: 1) the source node always chooses the best ingress overlay node with the least number of overlapping routers, 2) the best egress node to the destination is always selected, and 3) the path between the selected ingress and egress nodes does not experience any overlapping. It is possible that we undercount the number of shared routers of this best multi-hop path. However, this under-estimation is safe because we are using the multi-hop scenario as an upper bound comparison to the single-hop case.

The two lines in the graph in Figure 12(a) represent the path diversity of the single- and the optimal multi-hop overlay paths. We observe that a single-hop overlay path provides nearly the same degree of path diversity for most destinations as the optimal multi-hop overlay path does. This implies that when the direct Internet path fails, single-hop overlay backup paths are almost as reliable as multi-hop paths.

Next, we compare the latency of a single-hop overlay path with the latency of the optimal overlay path (i.e., shortest overlay path). The latency of a single-hop overlay path is computed by adding the roundtrip times of two paths: one between the source and an overlay node, and the other between the overlay node and a destination. For the optimal multi-hop overlay paths, we calculate the shortest roundtrip time for all overlay paths. For this calculation, we perform another measurement: we initiate ping probes among overlay nodes (i.e. 232*232 probes) in addition to probes between overlay nodes and endhosts. We assign this roundtrip time information as a weight to each virtual link between overlay nodes, and run SPF (Shortest Path First) calculations. Figure 12(b) shows the latency distribution of the direct Internet path, single-hop, and shortest multi-hop overlay paths. We observe that single-hop overlay paths improve latency for a large fraction of source/destination pairs compared with the direct Internet paths and perform as well as the shortest multi-hop paths.

Overall, we verify that in most cases single-hop overlay paths will suffice with respect to both path diversity and latency. This analysis result complements prior studies [13], [15]

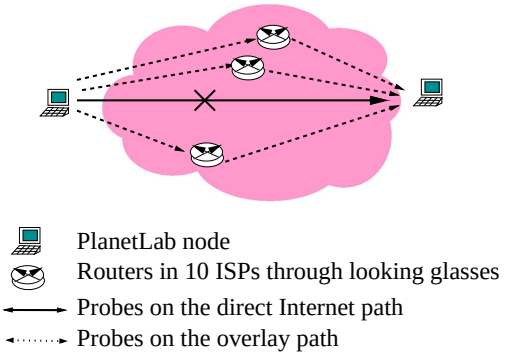


Fig. 13. Evaluation Scenario

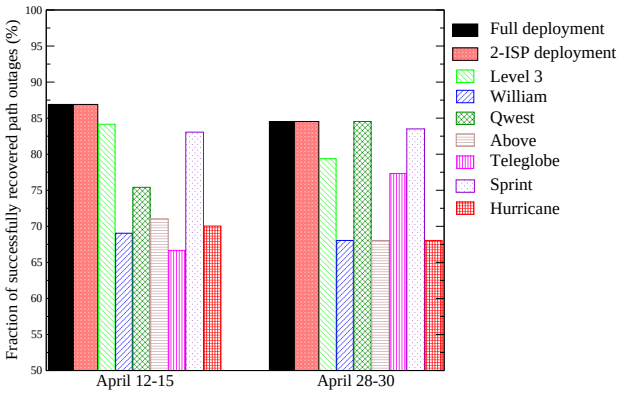


Fig. 14. Evaluation results: between ISPs

by providing measurement-based verification and insights to show that single-hop overlay routing performs as well as multi-hop routing in terms of both availability and performance. Recent work from University of Washington [33] confirms our observation. Taking a cue from this analysis, we adopt a simple but effective single-hop routing mechanism in our proposed topology-aware overlay framework. Single-hop routing does not require any complex routing mechanisms at the overlay layer. Overlay nodes in the single-hop routing approach can be considered as "relay" nodes which only forward the packets to destinations without making any routing decision. In contrast to existing overlay networks, these relay nodes do not need to exchange routing-related packets with each other. In addition, there is no extra delay from packets transiting more than one overlay node, making the single-hop routing solution more scalable.

Furthermore, we believe that by combining topology-aware node deployment, the single-hop routing mechanism becomes more effective in recovering from path outages and congestion problems. Since our node placement strategies ensure that each relay node is diverse from every other node, it is very likely that a proper relay node can be found which successfully provides a viable alternate path over the failed path. Also, source nodes can take advantage of the topology hints provided off-line to effectively select the most diverse relay node for each destination. We believe that this topology-aware single-hop routing is applicable to both reactive and multi-path overlay routing, enhancing availability and performance while reducing associated costs.

V. Evaluation

In this section, we evaluate the proposed topology-aware overlay framework by examining how well the proposed framework can recover from network failure events in the real-world. In particular, we statically choose a subset of ISPs and routers based on the clustering mechanisms combining path-diversity and latency, and we adopt the proposed single-hop routing. Note that the data sets used in Sections III and IV are snapshots of topology and latency, but they do not include logs of failure events. In contrast, the evaluation platform used in this section captures real-world failure events in real-time. With this platform, we are able to quantify how many network outages can be avoided by using our proposed framework.

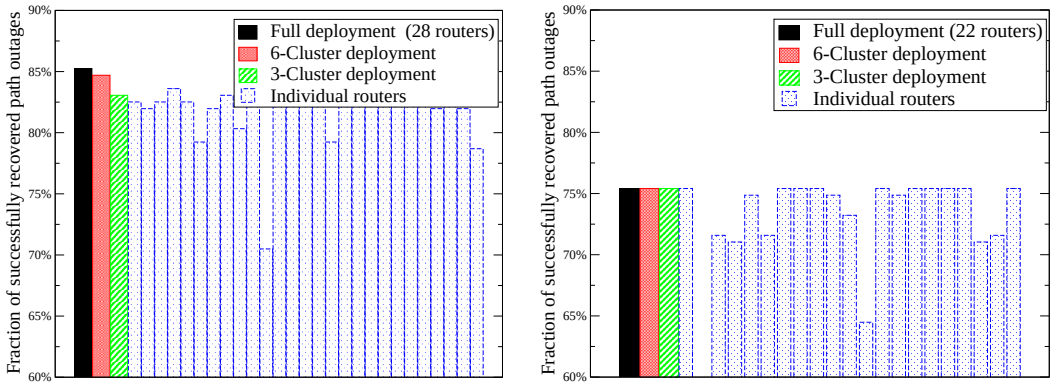
Our evaluation methodology is illustrated in Figure 13. We gather data on the end-to-end connectivity of direct Internet paths as well as overlay paths between PlanetLab nodes over two periods in April 2004. To monitor the direct Internet path, each PlanetLab node directly sends probes to randomly selected destinations at short intervals, without going through any intermediate overlay node. Each node independently sends ping probes and sleeps for a random period of time between 5 to 10 seconds.[4] When the probe fails, the node immediately re-sends at most three more probes at shorter intervals (less than 1 second). We define path failure as four consecutive lost probes (one original probe + three additional probes). When each node detects a path failure, it logs the observed outages of the direct path and also checks if the indirect overlay paths to the destination are available at that time. To check the availability of indirect paths, each PlanetLab node triggers two sets of probes: 1) from itself to overlay nodes and 2) from overlay nodes to destinations. To perform the second set of probes, we take advantage of ping utilities provided by looking glasses from 10 ISPs. After the probes of the overlay path are complete, the source node sends another probe along the direct Internet path to the same destination to make sure that the direct Internet path is still down. We consider a host to have failed if it stops sending probes for more than 10 minutes, and we discard probes lost due to host failure from our data sets.[5]

In Figure 14, we show the evaluation results at the ISP level. The left and right groups of bars represent two different experimental periods, April 12-15 and April 28-30, 2004, respectively. In each group, we present the results from full deployment, 2-ISP deployment, and individual deployment at seven ISPs.[6] For 2-ISP deployment, we use the same set

[4]Since we take advantage of public looking glasses, we put limitations on the probe rate to avoid overloading them.

[5]We exclude top 100 web sites in this evaluation because we are unable to check if the web site itself fails.

[6]We exclude three ISPs in the evaluation because the looking glasses of these ISPs were not stable during these periods.



(a) Sprint

(b) Qwest

Fig. 15. Evaluation results: between routers inside a single ISP

of ISPs throughout all time periods—we statically pick 2 ISPs based on measurements in Section III-C. Each bar on the graph represents the percentage of successfully recovered path outages. We observe that each individual ISP provides a different amount of failure recovery ranging from 65% to 83%. This implies that the choice of ISP can significantly impact the behavior of the overall system. This evaluation is consistent with our analysis in Section III-C in that any single ISP does not provide the best path diversity for every source and destination pair. On the other hand, by deploying overlay nodes at all 7 ISPs (i.e., full deployment), we can improve recovery to 87%. More interestingly, 2-ISP deployment provides the same degree of recovery as full deployment in both periods of experiments. This indicates that a carefully selected subset of ISPs can perform as well as full deployment throughout different time periods. Note that in the period of April 12-15, Level 3 contributes significantly to the full deployment. However, this does not necessarily mean that we only need to use Level 3 for constructing overlay networks. Consider the other experimental results from the April 28-30 data. In this period, a different ISP (i.e., Qwest) contributes significantly to the full deployment. This observation indicates that the best performing ISP changes over time depending on where link failures happen. Hence having all ISPs deployed (i.e., full deployment) would provide the best recovery performance. However, one of the goals in this paper is to maximize availability and performance while also minimizing deployment costs. In summary, the result in Figure 14 validates our argument that choosing a static set of k ISPs based on the proposed measurement-based clustering heuristics can achieve similar failure recovery as full deployment over all possible failures, regardless of time period.

In the next set of graphs shown, in Figure 15, we evaluate the overlay nodes inside a single ISP. Without loss of generality, we show the evaluation results for Sprint and Qwest during the period of April 12-15. In each graph, we present the results from full deployment, 6-cluster deployment, 3-cluster deployment, and individual deployment at each of 28 (or 22 in Qwest) routers. For 6- and 3-cluster deployment, we use the same set of routers throughout all time periods—we statically pick 6 (or 3) routers based on the measurements in Section III-B. We observe that each router provides a different amount of recovery. Also, we see that 3-cluster deployment provides almost the same degree of recovery as full deployment. Note that 3-cluster deployment in Sprint gives slightly worse performance than some well-behaving individual routers. This is explained by the fact that 3-cluster deployment does not include these well-behaving routers. While our static choice of three overlay nodes (i.e., 3-cluster deployment) does not include the best performing nodes in the study of Sprint, this 3-cluster deployment provides almost the same degree of resilience as full deployment. Overall, we believe that our algorithm for choosing overlay nodes provides a similar degree of resiliency as full deployment over all possible failures, regardless of time period. Recall that while a more dynamic choice of overlay nodes (full deployment) would provide better performance, we attempt to maximize resiliency while minimizing deployment costs.

There is another issue we should address. Many PlanetLab nodes are placed at Internet2 institutions which are connected to the Abilene backbone. Abilene is a backbone network available for communication between universities participating in Internet2. The direct Internet paths between Internet2 members travel the Abilene backbone. Hence, we suspect that most overlay nodes at commercial ISPs are likely to show a similar degree of outage recovery for link failures happening in the Abilene backbone because any non-Abilene overlay node can recover from an Abilene failure. If we extend our study to include more non-Internet2 members as destinations, we expect that each overlay node and ISP would show a wider range of recovery patterns.

Overall, these evaluation results show that the proposed approach is able to react and recover from about 87% of path outages, which is a significant amount of improvement.

Recall that existing overlay networks were unable to avoid about 50% of path outages [15]. Even if this study used a different evaluation platform, the amount of improvement by our proposed framework is still significant. Also, we observe that our proposed heuristics for choosing ISPs and overlay nodes provide almost the same degree of resilience as full deployment over all possible failures.

VI. Conclusion

Overlay networks are widely studied approaches, aiming to leverage the inherent redundancy of the Internet's underlying routing infrastructure to enhance end-to-end application performance and availability. However, the effectiveness of these overlay networks depends on the natural diversity of overlay paths.

In this paper, we presented a novel architecture for topology-aware overlay networks to maximize path diversity without degrading latency. First, we proposed several heuristics for overlay node placement based on analysis of extensive data collection from various vantage points. We showed that the proposed clustering-based deployment reduced the number of overlay nodes required but kept a high level of availability and performance. We believe that this off-line analysis study gives guidance to administrators to explicitly choose overlay nodes by considering underlying topology. Also, we provided measurement-based verification that a single-hop overlay path provided the same degree of path diversity as the multi-hop overlay path did for more than 90% of source and destination pairs. Applying single-hop routing in our framework, we are able to significantly reduce the overhead of extra routing at the overlay layer and also decrease delays in transiting overlay nodes. Finally, we showed that about 87% of path outages were avoided by the proposed approach in our real-world evaluation study. Although the proposed architecture targeted the application model with communications among "known" (or pre-registered) networks, we believe that this architecture can be easily extended for more general communications.

Overall, the results in this paper underline the importance of topology-awareness in overlay networks and direct future research for enhancing end-to-end application performance and availability. We also believe that the concept of topology-aware overlay networks is directly applicable to peer-to-peer applications and network security.

References

[1] C. Labovitz, G. Malan, and F. Jahanian, "Internet routing instability," *IEEE/ACM Transactions on Networking*, vol. 6, no. 5, pp. 515–528, 1998.

[2] C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental study of Internet stability and wide-area network failures," in *Proceedings of FTCS99*, 1999.

[3] C. Labovitz, A. Ahuja, and F. Jahanian, "Delayed Internet routing convergence," in *Proceedings of ACM SIGCOMM*, 2000.

[4] V. Paxson, "End-to-end routing behavior in the Internet," in *Proceedings of ACM SIGCOMM*, 1996.

[5] B. Chandra, M. Dahlin, L. Gao, and A. Nayate, "End-to-end WAN service availability," in *Proceedings of 3rd USITS, San Francisco, CA*, 2001.

[6] Z. M. Mao, R. Govindan, G. Varghese, and R. Katz, "Route flap damping exacerbates Internet routing convergence," in *Proceedings of ACM SIGCOMM*, 2002.

[7] A. Basu and J. G. Riecke, "Stability issues in OSPF," in *Proceedings of ACM SIGCOMM*, 2001.

[8] C. Alaettinoglu, V. Jacobson, and H. Yu, "Toward millisecond IGP convergence," in *Nanog 20*, Oct. 2000.

[9] A. Feldmann, O. Maennel, and Z. M. Mao, "Locating Internet routing instabilities," in *Proceedings of ACM SIGCOMM*, 2004.

[10] D. F. Chang, R. Govindan, and J. Heidemann, "The temporal and topological characteristics of BGP path changes," in *Proceedings of ICNP*, 2003.

[11] N. Feamster, D. G. Anderson, H. Balakrishnan, and M. F. Kaashoek, "Measuring the effects of Internet path faults on reactive routing," in *Proceedings of ACM SIGMETRICS*, 2003.

[12] J. Rexford, J. Wang, Z. Xiao, and Y. Zhang, "BGP routing stability of popular destinations," in *Proceedings of ACM IMW*, 2002.

[13] D. Anderson, H. Balakrishnan, F. Kaashoek, and R. Morris, "Resilient overlay networks," in *Proceedings of 18th ACM Symposium on Operating Systems Principles*, 2001.

[14] A. Collins, "The Detour framework for packet rerouting," M.S. thesis, University of Washington, 1998.

[15] D. G. Anderson, A. C. Snoeren, and H. Balakrishnan, "Best-path vs. multi-path overlay routing," in *Internet Measurement Conference*, 2003.

[16] J. Han and F. Jahanian, "Impact of path diversity on multi-homed and overlay networks," in *Proceedings of IEEE International Conference on Dependable Systems and Networks*, 2004.

[17] PlanetLab, "http://www.planet-lab.org," .

[18] H. Chang, S. Jamin, and W. Willinger, "Internet connectivity at the AS-level: An optimization-driven modeling approach," in *Proceedings of the ACM SIGCOMM workshop on models, methods and tools for reproducible network research*, 2003.

[19] skitter project, "http://www.caida.org/tools/measurement/skitter/," .

[20] R. Teixeira, K. Marzullo, S. Savage, and G. Voelker, "Characterizing and measuring path diverisy in Internet topologies," in *Proceedings of ACM SIGMETRICS (extended abstract)*, 2003.

[21] R. Braynard, D. Kostic, A. Rodriguesz, J. Chase, and A. Vahdat, "Opus: an overlay peer utility service," in *Proceedings of the 5th OPENARCH*, 2002.

[22] 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*, 1999.

[23] K. Nayak and D. McKernan, "Measuring provider path diversity from traceroute data: Work in progress," in *CAIDA-ISMA workshop*, 2001.

[24] A. Akella, B. Maggs, S. Seshan, A. Shaikh, and R. Sitaraman, "A measurement-based analysis of multihoming," in *Proceedings of ACM SIGCOMM*, 2003.

[25] W. Cui, I. Stoica, and R. H. Katz, "Backup path allocation based on a correlated link failure probability model in overlay networks," in *Proceedings of International Conference on Network Protocols (ICNP)*, 2002.

[26] A. Akella, J. Pang, B. Maggs, S. Seshan, and A. Shaikh, "A comparison of overlay routing and multihoming route control," in *Proceedings of ACM SIGCOMM*, 2004.

[27] S. Ratnasamy, M. Handley, R. Karp, and S. Shenke, "Topologically-aware overlay construction and server selection," in *Proceedings of IEEE INFOCOM*, 2002.

[28] M. Castro, P. Druschel, and Y. C. Hu, "Topology-aware routing in structured peer-to-peer overlay networks, Microsoft research technical report msr-tr-2002-82," .

[29] N. J. A. Harvey, J. Dunagan, M. B. Jones, and S. Saroiu, "SkipNet: A scalable overlay network with practical locality properties, Microsoft research technical report msr-tr-2002-92," .

[30] A. Nakao, L. Peterson, and A. Bavier, "A routing underlay for overlay networks," in *Proceedings of ACM SIGCOMM*, 2003.

[31] Z. Li and P. Mohapatra, "QRON: QoS-aware routing in overlay networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 1, pp. 29–40, 2004.

[32] J. Touch and S. Hotz, "The X-Bone," in *Proceedings of 3rd Global Internet Mini-Conference*, 1998, pp. 75–83.

[33] K. P. Gummadi, H. V. Madhyastha, S. D. Gribble, H. M. Levy, and D. Wetherall, "Improving the reliability of Internet paths with one-hop source routing," in *Proceedings of the 7th Symposium on Operating Systems Design and Implementation*, 2004.