# Testing the Scalability of Overlay Routing Infrastructures

Sushant Rewaskar and Jasleen Kaur

University of North Carolina at Chapel Hill,
{rewaskar,jasleen}@cs.unc.edu

**Abstract.** Recent studies have demonstrated the utility of alternate paths in improving connectivity of two end hosts. However studies that comprehensively evaluate the tradeoff between its effectiveness and overhead are lacking. In this paper we carefully characterize and evaluate the trade-off between (1) the efficacies of alternate path routing in improving end-to-end delay and loss, and (2) the overheads introduced by alternate routing methodology. This would help us to test the scalability of an overlay network.
We collected ping data on *PlanetLab* and studied the above trade-off under different parameter settings such as path sampling frequency, overlay-connectivity, number of overlay hops etc. Results from ten sets of measurements using 35 to 36 of the *PlanetLab* nodes are used to test the effect of the various parameters. We find that changing epoch duration and connectivity helps reduce the overheads by a factor of 4 and 2 respectively at the cost of some lost performance gain.

## 1 Introduction

Recently, the idea of using overlay networks to find routes between a pair of Internet hosts has received much attention [1–3]. In these works, hosts associate with the "nearest" node belonging to an overlay network, and route traffic through that node. The overlay node forwards traffic to the specified destination along the current best overlay path in such a manner as to provide better end-to-end services such as delay, loss, jitter, and throughput. In order to find the best path, overlay nodes will: (i) monitor the state of routes to all overlay nodes, (ii) periodically exchange this state information with neighboring overlay nodes, and (iii) compute an up-to-date snapshot of the best paths between any pair of overlay nodes using a link-state routing protocol. Clearly, the gains in end-to-end performance achieved through the use of an overlay network come at the cost of processing and transmission overhead (overlay state monitoring and distribution, and path computation). Past work has shown that it is possible to construct an overlay mesh of up to 50 nodes without incurring significant overheads [2]. In this study, we ask the question: *can one scale an overlay up to a larger number of nodes?*

The fundamental issue is that the tradeoff between the performance gains and overheads of an overlay routing infrastructure is governed by several factors, including (i) the number of nodes in the overlay, (ii) the average fraction

of "neighbors" of a node in the overlay with whom a node exchanges routing information (referred to as *logical connectivity* or simply *connectivity*), (iii) the frequency with which routes in the overlay network ("overlay links") are computed (the duration between route updates is referred to as an *epoch*), and (iv) the maximum number of overlay "hops" used in computing routes. Past work has primarily investigated single points in this parameter space in wide-area Internet measurements. Here we attempt a more comprehensive analysis of the limits and costs of the scalability of an overlay routing infrastructures by a more controlled study of the parameter space. Our specific goal is to identify points in the parameter space that would enable the operation of larger overlay routing infrastructures with acceptable overheads and satisfactory performance gains.

We have conducted an extensive measurement and simulation study of a wide-area overlay routing infrastructure using PlanetLab. Based on this study our main findings are:

1. Reducing the average logical connectivity of overlay nodes by a factor of 2, reduced overlay routing maintenance and messaging overhead by almost a factor of 4, while reducing by only 40% the number of overlay routes having better latency than the default path, and reducing by only 30% the number of overlay routes having lower loss rates than the default route.
2. Doubling epoch duration reduces overhead by 50%. However, as epoch duration increases, routing data becomes more stale (less valid) and the performance of the overlay routes computed based on this data is less certain. For large epoch duration, computed overlay routes underperform nearly 30% of the time for loss rates and nearly 10% of the time for latency. The selection of wrong routes due to stale information is termed as mis-predictions.
3. Increasing the number of nodes in the overlay by 33% increases the number of better paths by 5 to 10% while increasing the overhead by about 60% in the worst case (when nodes have 100% logical connectivity).
4. Using more then one intermediate path for calculating an overlay route does not provide much benefit.

The rest of this paper is organized as follows. In section 2 we place our work in context of other related work, highlighting the similarity and differences. Section 3 describes the experimental and analysis methodology. The dataset is described in brief in section 4. We present the results in section 5. We summaries our results and describe scope for future work in Section 6.

## 2 Related Work

Overlay networks provide a framework for building application-layer services on top of the existing best-effort delivery service of the Internet. Key to this work is the understanding of the dependence of end-to-end performance on route selection on the Internet. This problem has been studied extensively. For example, the detour[5] study quantifies the inefficiencies present in the Internet direct default paths. They collected data over a period of 35 days over 43 nodes. Based

on this data they concluded that overlay paths can improve network performance. The Savage *et al.*[3] study conducts a measurement-based study where they have compared the improvement in performance that can result from the use of an alternate path instead of the direct default path. They studied various metrics such as loss and round trip delay over a set of diverse Internet hosts. They conclude that alternate paths could be helpful in 30 to 80% of the cases. Most recently Anderson *et al.* [2] showed the use of an overlay network to achieve quick recovery in case of network failures. Unlike detour, the RON work analyzed the use of alternate paths for improving the performance on a time scale small enough to reflect the actual dynamic changes in the Internet.

All these papers considered only a single point in the parameter space to study system performance. None studied the effect of varying the parameters and their effects on the various metrics. For example, based on a single point in the parameter space, [2] predicts the size to which such a network could be scaled.

## 3 Overlay Network Emulation Facility

### 3.1 Network Node Architecture

We emulate an overlay network by gathering ping data from a collection of nodes in a real wide-area network and then subsampling the data as appropriate to emulate an overlay network with the desired degree of logical connectivity between nodes. Each node in the network runs two programs: a "ping" module that emulates the probing mechanism in an overlay and a "state exchange" module that emulates mechanism used to propagate measurement data to other nodes.

At the beginning of each epoch, the ping module collects ping measurements to all other nodes. It then summarizes this ping information and records it in a local file for off-line analysis. Following each 24 or 48 hour measurement period, data from the nodes is collected and analyzed. To emulate different average connectivity of nodes, the analysis considers only information that would be acquired about neighboring nodes. The analysis tool uses the ping summaries to compute, for each service metric such as delay and loss, all "better" overlay routes to all other overlay nodes. Only those alternate paths are considered that use one of the neighbors as the next-hop. We also ran the ping module on each node and used a link state routing protocol to flood the summary of the measurements to all nodes. The bytes exchanged by this module was used to determine the state-exchange overhead.

### 3.2 Parameters

The tradeoff between performance gains and overheads of a routing overlay is governed by several factors:

**Epoch** : This is the interval at which the probes are run and routing updates are conducted. The larger is the epoch duration, the less frequent are route computations, and smaller are the overheads. On the other hand, smaller epochs help maintain an up-to-date view of network state and best routes.

**Average connectivity** : The larger is the set of neighbors used to re-route traffic, the greater is the likelihood of finding better alternate paths. However, the overhead of exchanging state and computing best paths also increases with connectivity.

**Overlay size** : The larger is the number of nodes in the overlay, the greater is the likelihood of finding better alternate paths. However, the overheads grow as well. In fact, our main objective in this study is to find out if the other parameters can be tuned to allow the operation of larger overlay.

**Maximum length of overlay routes** : We expect that considering only routes that traverse no more than 2-3 overlay links will be sufficient for locating better paths, if there are any.

**Set of service metrics** : The likelihood of finding better alternate paths between a pair of overlay nodes is a function of the service metrics - such as delay, loss, and jitter - of interest. Furthermore, applications that desire good performance in terms of more than one metric stand a smaller chance of finding paths that do better than the direct paths. The overheads remain unaffected by the set of service metrics.

These are the set of parameters what can be tuned to increase or decrease the performance gains and overheads of the overlay for desired results.

### 3.3 Metrics

We test the scalability limits of overlay routing infrastructures by conducting several experiments with different settings of the parameters mentioned in previous section and measuring the gains and overheads from each. Gains and overheads are quantified as follows.

**Gain Metrics** Four metrics are used to quantify the performance benefits achievable with overlay routing:

1. The number of node-pairs for which there is at least one alternate path that is better than the direct path.
2. The number of better alternate paths for a node pair.
3. The degree of performance improvement achieved by using the best alternate path.
4. The number of mis-predictions that occur due to stale state information, when large epoch durations are used.

**Overhead Metrics** We measure overheads in two different ways:(i) the ping overhead, and (ii) the state exchange overhead, computed in terms of the bits introduced into the network per second. For really huge overlay networks the computational cycles required to calculate a alternate path will also become significant but are not covered here.

## 4 Data Collection

**Table 1.** Characteristics of the Ping datasets used in this paper.

| dataset | Dates | Duration | No.of nodes |
|---|---|---|---|
| D1 | 2-3 Feb,04 | 24 Hours | 36 |
| D2 | 3-4 Feb,04 | 24 Hours | 36 |
| D3 | 5-6 Feb,04 | 24 Hours | 36 |
| D4 | 6-7 Feb,04 | 24 Hours | 36 |
| D5 | 7-9 Feb,04 | 48 Hours | 35 |

**Table 2.** Dataset for the state exchange overhead information

| dataset | Dates | Connectivity | No.of nodes |
|---|---|---|---|
| D61 | 17-18 Aug,03 | 100% | 36 |
| D7 | 8-9 Sep,03 | 50% | 34 |
| D8 | 3-4 Sep,03 | 25% | 36 |
| D9 | 10-11 Sep,03 | 12% | 34 |
| D10 | 11-12 Sep,03 | 6% | 34 |

Our experiments were performed on the PlanetLab testbed [4]. PlanetLab is an open, globally distributed testbed for developing, deploying and accessing planetary-scale network services. The PlanetLab testbed consists of sites on both commercial ISPs and the Internet2 network. We selected 36 nodes all over the North American continent. We did not select sites outside the North American continent as many alternate path to these sites would, in most case, share the same transoceanic link and would have provided less chance for performance gain. For 100% connectivity a total of up to (36*35=1260) paths are monitored.

To analyze the gains we collected five datasets as described in Table 1. Each data set was then analyzed offline with 100%, 75%, 50%, 25% and 12% connectivity. The neighbors of each node were selected at random. The dataset was also used to estimate the effect of increasing epoch duration. Analysis were done on the first dataset to study the effect of reduced number of nodes in the overlay. Performance for 9, 18 and 27 randomly selected nodes was studied.

We also studied the overhead of routing with the flooding protocol implemented in our system (number of bytes received and send by each node), as a function of logical connectivity. Table 2 describes this data.

## 5 Planetlab Results

### 5.1 Effect on Gain

Here we present the effect of varying parameters on the gains achieved by the overlay network.

**Connectivity** We analyzed the dataset for different degree of average node connectivity. Connectivity is the total number of nodes about which a node has information. It could have information about a remote node because it ping at it or because it has been pinged by the remote node and received the information

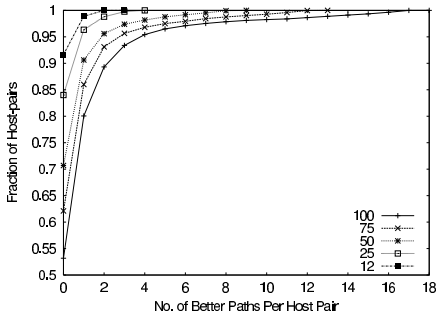


**Fig. 1.** Number of Better alternate paths for Latency

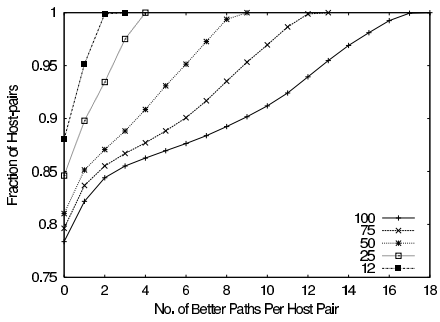


**Fig. 2.** Number of Better alternate paths for Loss

from the remote node. Thus to achieve 100% connectivity a node need to ping only 50% of the nodes as long as all the remaining 50% of the nodes ping it.

Figure 1. shows the number of better alternate paths per host-pair for latency. We can see that reducing the connectivity reduces the number of host-pairs for which better paths can be found, however, there are still significant number of host-pairs with better alternate paths. For e.g. for 100% connectivity 53% of host-pairs had no better path i.e. 47% of the paths had at least one better path. Reducing the connectivity to 75% increased the number of host-pairs with no better paths to 61% i.e. still over 39% of the paths had a better alternate path. Here decrease in the number of host-pairs having better alternate paths is around 17%. It is also seen that in many cases there exists more then one better alternate paths. This suggests that even after removing a few nodes from the neighborhood of a node it should be able to find at least one better path.

Figure 2. shows a similar plot for loss. Reducing connectivity to 75% still allows us to find a better path for 20.3% of the host-pairs against 22% at 100% connectivity. The decrease in the number of host-pairs with better alternate paths is only around 2%.

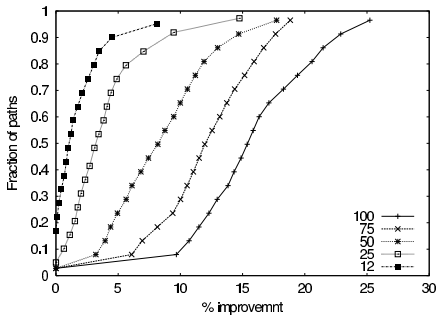


**Fig. 3.** % Improvement for Latency

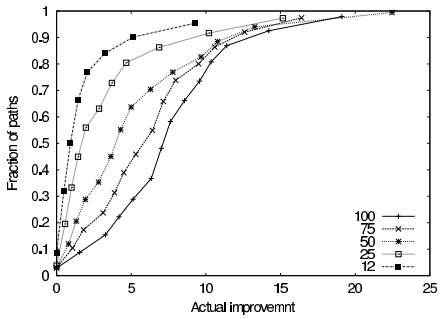


**Fig. 4.** Actual Improvement for loss

Figure 3. and 4. shows the improvement that is achieved for latency and loss respectively. For latency percentage improvement is plotted. However, for loss in many cases the percentage improvement is 100% (loss changing from a high value to zero) and hence the actual improvement is plotted. When connectivity is reduced from 100% to 75% the median of the improvement achieved for latency goes down from 15% to around 12% and for loss it goes down from 7% to 6% .

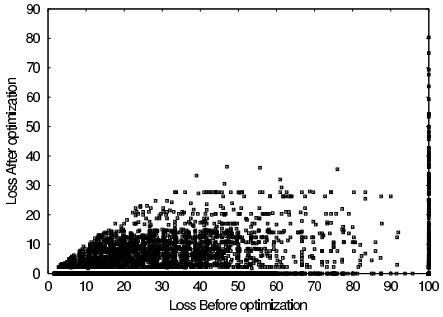


**Fig. 5.** Actual Improvement for loss

For Protocols like TCP an improvement of 3% in the loss is more significant when the change is from 3.5 to 0.5% rather then 93 to 90%. In Fig 5. we see that in most of the cases the improvement in loss is significant in these terms too i.e. the loss on the best path is very close to zero. We also observed that this type of improvement is independent of the connectivity or epoch duration for the overlay.

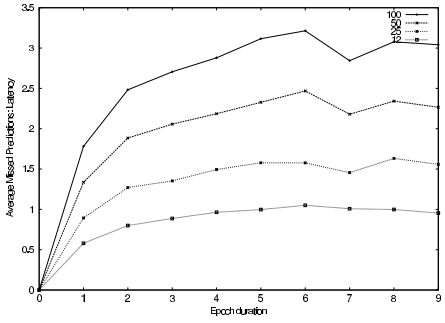


**Fig. 6.** Number of Mis-prediction as a fraction of total predictions for Latency

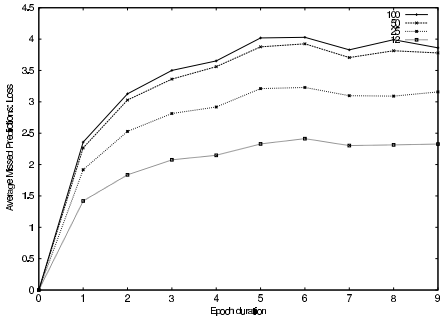


**Fig. 7.** Number of Mis-prediction as a fraction of total predictions for Loss

**Epoch duration** Figure 6. shows the number of erroneous best path predictions for latency resulting from an increase in the epoch duration keeping the connectivity constant. The fraction of mis-predictions with a larger epoch is computed relative to the epoch duration used in the measurements. The X-axis of the graph indicates the epoch duration and the Y-axis shows the number of mis-predictions as a fraction of the total better alternate path computed. By just doubling the epoch duration the number of mis-predictions may go up by around 10% at 100% connectivity. Varying connectivity from 100% to 75% increase the number of mis-predictions by around 2%.

Figure 7. shows a similar graph for the number of erroneous predictions for loss. Doubling the epoch duration results in around 30% mis-predictions at 100% connectivity. Here the number of mis-predictions seems to be less affected by connectivity.

### 5.2 Overlay size, Number of Hops and Service Metrics

In this section we discuss the effect of the number of nodes in the overlay, number of intermediate hops and service metrics.

It is seen that increasing the overlay size aids in identifying more number of better path for both loss and latency. This is expected, as we increase the number of nodes the number of paths scanned for a better path increases and hence the probability of finding a better path increases. The amount of improvement that can be achieved also increases. We find that increasing the overlay size by 33% increases the number of better path identified by around 5 to 10% and the improvement on these paths by around 5%. The number of mis-predictions does not seem to be affected by the scale of the network. This however would increase the overheads by about 60% for 100% connectivity.

We studied the effect of using multiple nodes to find a better alternate path for a given host-pair (i.e. instead of going through a single intermediate node we use 2 or more intermediate nodes). We found that multiple hops do not contribute significantly in improving the performance seen by a given node.

We have conducted the experiments to identify better path with respect to loss and latency. If application requires better paths for different metric or multiple metrics the performance gain may differ. For example, if we need to find better path for three metrics (loss, latency and jitter) simultaneously, the total number of better alternate paths is just 1%. If we need to find better paths for only loss and jitter about 7 to 8% of the host-pairs where seen to have a better alternate path.

### 5.3 Overheads

For computing better paths, an overlay needs to collect and distribute the network state. The process of sampling the network and distributing the state causes extra traffic on the network (overheads).There are two types of overhead associated with our method:

– Ping: Ping is used to collect current information about loss and delay on the network. Each node is pinged for maximum of six seconds with the amortized rate of pinging being 6 packets per second. In our experimental set up with 100% connectivity and 36 nodes, ping introduced a traffic of 3Kbps during the active period (when we are sampling the network). Now the actual epoch length consists of the active periods and the passive periods (when no sampling is done) hence the total network traffic generated over a epoch is the number of bits sent divided by the epoch length and would be even smaller then 3 Kbps.
– State Exchange: Once the nodes collect information regarding other nodes it is connected to, we use a link state protocol to transmit this information to all nodes. The overhead in this is the numbers of bytes that have to be sent and received by each node.

We analyze the effect of varying parameters on both types of overheads

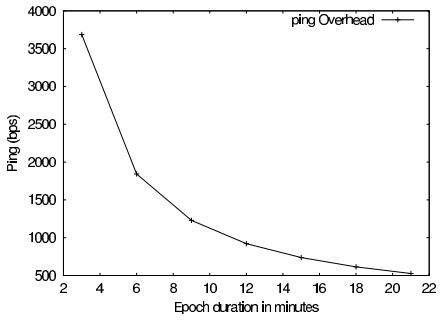


**Fig. 8.** Ping Overheads

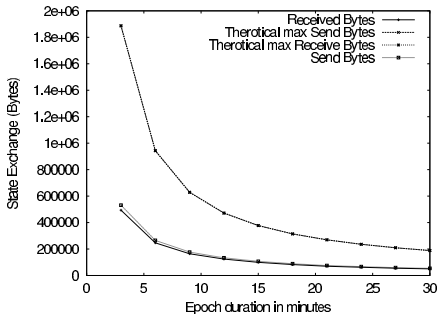


**Fig. 9.** State exchange overhead

**Ping** Figure 8 shows the effect of increasing epoch duration on ping overhead. As can be seen as the epoch duration increases the average amount of bytes sent in the network for active measurements decreases. To estimate whether an alternate path is better then a direct path we need to have information regarding the direct path between all host-pairs. Hence even for reduced connectivity we ping all the nodes. Connectivity only affects the amount of data exchanged and not the ping overhead.

Figure 9. shows the effect of epoch duration on the number of bytes sent and received by Kupl1.ittc.ku.edu with all other nodes. The theoretical maximum amount of bytes exchanged corresponds to the amount of bytes that would have been exchanged if we exchange the complete information gathered at any epoch. However in our setup we exchange only the information that has changed since the last epoch and hence the total amount of information exchanged is very low. Changing epoch duration obviously reduces the amount of data transmitted.

Reducing the connectivity also affect the amount of data exchanged. Reducing connectivity from 100% to 50% reduces the amount of data transmitted by almost a factor of 4 from 11kbps to 3kbps.

Ping as well as the state exchange overhead increases with the increase in the overlay size. Doubling the overlay size increases the amount of overhead bytes by about 4 times.

## 6 Future Work and conclusion

This paper presents an overview of the tradeoff between performance gain in an overlay network with the overheads incurred. Studying these we can tune the overlay to achieve the required performance while controlling the overheads.

Based on the analysis using actual network measurements we conclude that alternate paths help improve loss and delay over a network. Moreover, connectivity influences these improvements. Reducing connectivity by half reduces the overhead by four times at the cost of reducing the number of better alternate paths by almost 40% for latency and over 30% for loss. Similarly increasing the epoch duration by two reduces the overheads by two. However, this may lead to mis-predictions in computing the best paths for loss in almost 30% of the cases, and for latency in 10% of the cases. As the epoch duration increases the number of mis-predictions increases but soon stabilizes, however the maximum number of mis-prediction may be quite high in some cases. Thus by reducing the connectivity by half and increasing the epoch duration by four we can achieve a network 8 times the current size with the same amount of overhead but some lost performance.

There are several issues that warrant further investigation. We want to investigate the possibility of using passive sampling to gather network conditions instead of active sampling of the network. The effect of adding nodes outside the North American continent also needs to be studied. Also we need to calculate the computational overhead incurred for path calculation at any node.

## References


1. D. Andersen, A. Snoeren, and H.Balakrishna: Best-Path vs. Multi-Path Overlay Routing. Proc. of the ACM SIGCOMM Internet Measurement Conference. Miami, FL, October 2003.
2. D. Andersen, H.Balakrishna, F Kaashoek, and R. Morris: Resilient Overlay Networks. Proc. of the 18th ACM Symposium on Operating System Principles, October 2001.
3. S. Savage, A. Collins. E. Hoffman, J. Snell and T. Anderson: The End-to-end Effects of Internet Path Selection. Proc. of ACM SIGCOMM, September 1999.
4. Peterson, L., Anderson, T., Culler, D., Roscoe, T : A Blueprint for Introducing Disruptive Technology into the Internet. Proc. of ACM HOTNET, Oct. 2002
5. Savage,S.,Anderson,T., Et Al., Detour:A Case for Informed Internet Routing and Transport. IEEE Micro 19, 1 (Jan. 1999), 50-59.