Informed Detour Selection Helps Reliability

Boulat A. Bash
Computer Science Department
University of Massachusetts
Amherst, Massachusetts 01003
Email: boulat@cs.umass.edu

***Abstract*—In this work we propose to use path information to improve the reliability of the Internet. Previous work put forth a simple idea of using an *overlay* network of intermediary detour nodes that can be used to route around failures on direct Internet paths [1]. We provide the mechanism for choosing these intermediary nodes.**

The underlying idea of our work is that knowledge of the point-of-presence (PoP) path between the source and destination as well as the PoP paths between potential detour nodes and destination can be used to pick the intermediary for routing around failures on the direct path. We validate our proposal by performing an experiment using PlanetLab [2] and leveraging the existing iPlane [3] infrastructure. Using the data from iPlane, we obtain a two-fold decrease in path outage probability over the state-of-the-art.

I. Introduction

As the Internet continues to evolve, reliability remains a key issue. Research has shown that the Internet in its current form fails to achieve the “five nines” (99.999%) of connection reliability that the public switched telephone network (PSTN) demonstrates [4], instead realizing between 96.7% and 98.5% reliability according to various studies [5]–[7]. There are a number of reasons for this: server failures, router misconfigurations, and long BGP response times. One could argue that this is to be expected of a network that provides a best-effort datagram service where most of the intelligence lies at the end-hosts (as opposed to the network itself, like in PSTN). However, 98.5% availability is insufficient for many applications, including medical collaboration and some financial systems. Currently, these applications utilize expensive, dedicated networks to assure the required availability. In this paper, we propose an architecture that improves the reliability of the Internet paths and report the results of the validation that show a two-fold decrease in path outage probability over the state-of-the-art, achieving 99.2%–99.5% path availability.

There are two main approaches for making the Internet service more reliable: adding server redundancy and path redundancy. Note that these methods are complimentary; if one adds redundant servers in topologically distinct regions of the Internet, one also increases path redundancy. Server redundancy is mainly achieved using content distribution networks (CDNs) [8]–[10]. While these methods are popular, they unfortunately benefit only a few kinds of traffic, such as web page fetches.

Path redundancy can be applied to mitigate the Internet path failures. The idea is to navigate around the failed portion of the path using one or more *detour* nodes [11]. Utilizing such path redundancy for reliability was first implemented in Resilient Overlay Network (RON) system [1] by Anderson *et al.* RON used aggressive monitoring of a relatively small overlay network (16 nodes) to recover from faults, and identify and exploit opportunities for performance improvement (such as utilizing violations of the triangle inequality to decrease latency). Unfortunately, wide scale deployment of RON requires significant monitoring overhead, since the overlay continually probes the complete graph between all of its nodes.

A light-weight detouring method that does not require path monitoring was proposed by Gummadi *et al.* and is called Scalable One-hop Source Routing (SOSR) [12]. They demonstrate that it is usually sufficient for failure recovery to attempt to route indirectly using 4 *randomly chosen* nodes from a sufficiently geographically distributed set of candidate detour nodes (they used 67 PlanetLab [2] machines as potential detour node set.) If one of the detour paths succeeds, then fault is avoided. Effectively, they utilize the idea of having multiple random choices used in load-balancing [13]. They present a Linux implementation and show that they can route around 56% of network failures.

We propose to combine the lightweight SOSR approach with the idea of *informed* (as opposed to random) choices of the detour nodes as done in RON. However, unlike RON, our system would not aggressively monitor the links in the Internet. The knowledge we require in order to make an informed decision on which detours to try is the point-of-presence (PoP) path between each candidate for the detour node and the destination. Point-of-presence is a collection of routers belonging to the same autonomous system (AS) that are close enough topologically that they have similar routing tables (usually, points-of-presence are inferred from geographical locations of routers within AS). Because earlier studies [6] have shown that the Internet paths are stationary over the course of 24 hours, it is sufficient for the detour-to-destination paths to be updated daily.

Since knowledge of the detour-to-destination paths is not free, our proposed system is not as scalable as SOSR. We are effectively trading off the scalability for increased path availability. In this study, we validate our proposal by exploiting an existing and unrelated system for these data: a PlanetLab-based Internet map project called iPlane [3]. iPlane provides two valuable services: IP prefix to PoP mapping and daily traceroute data from most PlanetLab sites to destinations in almost every PoP. Like in SOSR, we use PlanetLab machines as our detour nodes. However, our client ranks the detour nodes by how much the PoP path (i.e. path according to the most current traceroute with IP addresses mapped to PoP IDs) from the detour node to destination overlaps with the PoP path from the client to the destination. Less overlap is better. In this work we test two metrics for path overlap: a count of common PoP IDs on two paths, and a count of common PoP links on two paths. As we show in Section III, both yield almost identical results. We select the top k nodes as detour intermediaries, where $k = 1, 2, 3, 4, 5$ in this work (but k can potentially be larger.)

In the next section we describe the methodology behind the experimental validation of our proposal, and in Section III we demonstrate that our proposal significantly improves reliability over SOSR. In Section IV we discuss the reasons for this improvement. We examine the related work in Section V and conclude with the ideas for future projects in Section VI.

II. Methodology

In order to compare the performance of our informed detour selection to random selection, we implemented a distributed Internet measurement system on PlanetLab [2]. This section describes the methods we used and our system design.

We used 50 PlanetLab nodes as the vantage points for our experiment, which were selected randomly from the set of 121 PlanetLab machines that had a high-bandwidth connection to the Internet and were reachable at the start of the experiment. Each vantage point probed a subset of destinations randomly selected from a reachable subset of random routers on the Internet (random .1 IP addresses). We restricted ourselves to routers and did not include end-hosts because we are mainly interested in the availability of *paths*. Routers are less likely to fail than end-hosts, and are more likely to consistently return probes. The destination lists were disjoint across our vantage points, thus, during normal operation, each destination was probed by one PlanetLab node.

The probing was done by pinging each destination every 15 seconds. The path was considered failed if two consecutive pings were missed. When the path failed, the vantage point responsible for probing the destination requested that PlanetLab nodes on its *intermediary set* ping the destination of the failed path and answer if the ping is successful. These requests were sent out every 15 seconds while the path to the destination was down. The intermediary set for each vantage point was the set of all 121 PlanetLab nodes we identified as usable excluding the vantage point and nodes on its site (for example, the Harvard site contains two PlanetLab nodes, neither of which would be included in either intermediary set) and containing only one randomly selected node per site (thus, none of our PlanetLab vantage points used both Harvard nodes as intermediaries). This was done in order to minimize the load on PlanetLab nodes, as well as reduce the traffic generated by the experiment. The set of intermediaries was the set of potential detour nodes that a client in a vantage point could use in case of a path failure.

We used two metrics to pick intermediaries for detouring: a count of common PoP IDs and a count of common PoP links on the paths from original vantage point and intermediary to the destination. When determining links, we removed the unknown hops returned by traceroute (“0.0.0.0” hops) as well as those hops that were not present in the iPlane IP-to-PoP mapping and used hops with known PoP IDs as link ends. Thus, if a traceroute contained the following path: “IP in PoP-1, unknown IP, unknown IP, IP in PoP-2, IP in PoP-3”, our system used $\{$PoP-1, PoP-2$\}$ and $\{$PoP-2, PoP-3$\}$ as links. We also did not use unknown hops in the count of common PoP IDs. However, we did record unknown hops in the path hop count, which was used as the tie-breaker in our metric (we preferred shorter paths.)

We ran our experiment for 379 hours from 6:00 AM EST March 25th 2008 until 1:00 AM EST April 10th 2008. While the experiment was running, we downloaded the daily traceroute files from the iPlane website

[14]. We were only able to use data from 46 vantage points, since some PlanetLab nodes were reset or rebooted without yielding data. 35 of those vantage points were located in the United States, 8 were in Europe, and 3 were in Japan. 42 vantage points were hosted by various universities and the rest by industrial research labs. Each vantage point had access to between 83 and 97 intermediaries, of which between 74 and 90 were hosted by universities. For any given vantage point, between 58 and 71 intermediaries were located in the United States, 15 to 20 in Europe, and between 4 and 7 in Asia. Thus, the distribution of vantage points across both the academic affiliation and geographic domains roughly corresponded to the distribution of the intermediates.

III. Experimental Results

Our system recorded $54,793$ path outage events, with mean and median outage durations of $1,309.41$ and 120 seconds, respectively. In aggregate, the path was in outage 1.477% of the time. The duration of outages that we witnessed had a heavy-tailed distribution, as evidenced by the scatter plot of the empirical complimentary cumulative distribution of the observed outage durations on Figure 1. This is consistent with the literature [5], [15]. There were some paths that were down for a long time: the longest outage we detected lasted 508,905 seconds, or 6 days 17 hours 21 minutes and 45 seconds.[1]

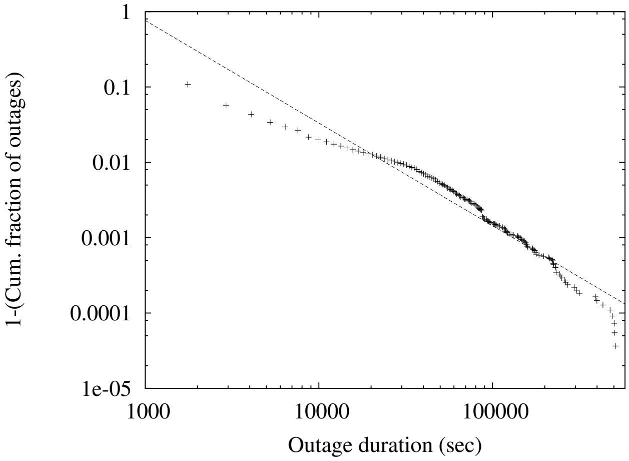

Fig. 1. Empirical complimentary cumulative distribution of outage duration on $\log_{10}$-$\log_{10}$ scale, showing that it is heavy-tailed.

Our experimental framework records the intermediaries that were successful in reaching the destination. Thus, we could mimic a system where intermediaries are detour nodes used during path failures. Alternate path to a destination through at least one of the intermediaries was available during $44,276$ out of $54,793$ outages (80.8%). The mean number of intermediaries with an available path was 9.1, the median 8. We can see from the histogram on Figure 2 that the distribution of the number of intermediate paths looks somewhat bimodal when paths exist, with peaks around 2 and 16. The reasons for this are unclear. The paths through an intermediary whose path had the least number of common PoP IDs and links with direct path were available $12,646$ times and $12,642$ times (both $\approx 23.1\%$), respectively. Table I summarizes the availability of paths through intermediaries.

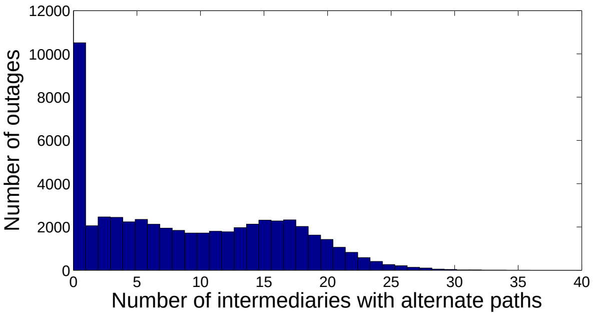

Fig. 2. Histogram of path availability across outages

TABLE I
Availability of alternate paths

Number of outages:	$54,793$
Number of outages with alternate path:	$44,276$
Mean num. of intermediaries with alt. path per outage:	9.1
Median num. of intermediaries with alt. path per outage:	8

k	By common PoP count	By common link count
1	12,646 (23.1%)	12,642 (23.1%)
2	17,357 (31.7%)	17,231 (31.4%)
3	21,322 (38.9%)	21,449 (39.1%)
4	24,550 (44.8%)	24,568 (44.8%)
5	24,550 (44.8%)	24,568 (44.8%)

(k is the number of top intermediaries ranked by corresponding metric. E.g. in $21,322$ of total $54,793$ outages (or 38.9%) at least one among the top three intermediaries ranked in the increasing order of path overlap via PoP count metric had working path to destination.)

Now let us examine the performance of the SOSR-like system described in the introduction. Recall that the original SOSR system attempts to use 4 randomly-chosen detour nodes to recover from faults [12]. We

[1] One must note that there were 26 intermediaries that we could have used to route around the failure, including one whose path had the least number of common links with the direct path.

select from one to five intermediaries in case of an outage and attempt to continue connection. Only if none of the intermediaries can reach the destination does the connection remain in outage. Otherwise, the system succeeds in preventing the break in the connection.

The probability that the random selection of k intermediaries fails in the case of n potential intermediaries and $m \leq n$ intermediaries that have a working detour path to the destination can be expressed as follows:

$$\mathbf{P}\left(\text{failure using } k\right) = \frac{\binom{n-m}{k}}{\binom{n}{k}} = \qquad (1)$$

$$= \prod_{i=0}^{k-1} \frac{n-m-i}{n-i} \qquad (2)$$

We use the equation (2) in our programs to compute the outage probability using SOSR-like random-k method.

Table II illustrates that our informed selection methods substantially outperform the random intermediary selection. Outage probability is computed by taking the total time in outage across all paths, subtracting the time in outage (expected time in case of SOSR-like random-k method) where the detour exists (given the selection method), and dividing by total time paths were monitored. Note that the path availability for each method can be obtained by taking the complement of the corresponding outage probability given on Table II. We are able to achieve “two nines” of reliability with either method and selecting just one intermediary node, while the random selection fell short even utilizing five detour nodes. We also note that the two informed methods are not substantially different in performance.

TABLE II
IMPACT OF INTERMEDIARY SELECTION METHODS ON PATH OUTAGE

k	Outage probability by selection method			
	Random (SOSR)	Common PoP count	Common link count	Geographic random
0	1.477%	1.477%	1.477%	1.477%
1	1.295%	0.757%	0.743%	1.261%
2	1.147%	0.648%	0.654%	1.089%
3	1.024%	0.566%	0.571%	0.952%
4	1.023%	0.516%	0.527%	0.841%
5	1.022%	0.516%	0.526%	0.750%

(k is the number of intermediaries used)

Our system monitored a total of 4,269 paths. 1,715 (40.2%) of those paths did not experience a failure that we detected, and 328 (7.7%) paths had failures but no intermediary with an available path to the destination (we suspect that these were due to destination failures as opposed to path failures). Examination of the plot of the cumulative fraction of paths vs. their availability on Figure 3 reveals that the informed method achieves considerable availability gains for the paths that are unavailable for substantial periods of time. In many cases of long-duration failures, the informed method was able to find a working alternate path where a random method would have had a high probability of failure due to the comparatively small fraction of intermediaries having working paths to those destinations.

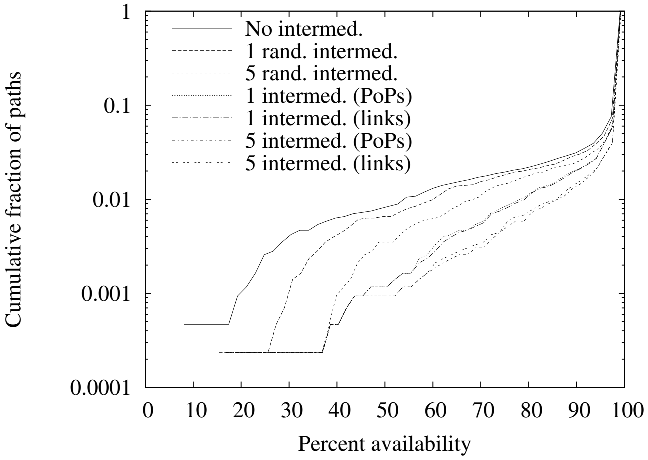

Fig. 3. Cumulative fraction of paths vs. percent availability. Note that the curves corresponding to the informed methods are substantially to the right of the curves corresponding to random selection. This illustrates that the availability gain for the informed selection is on average greater in magnitude then of the gain for random selection.

IV. Discussion

The premise of the SOSR work is that the paths between various detour nodes and the destination are independent enough that random selection works well. Our results put this presumption into question, since informed selection should help only when there is significant systemic correlation between the path from the vantage point to the destination and a significant subset of the paths through the potential detour nodes. We identify two sources of such correlation in our study:

- The unique feature of PlanetLab design, with many of its nodes belonging to the universities, since, during a fault in Internet2’s peering to a certain destination, a large subset of the potential intermediaries in the universities may experience correlated connectivity failures; and

- Geographic dependencies, as geographically close nodes are more likely to have correlated paths, thus leading to correlated failures.

While recording the structure of the paths between the vantage points, intermediaries and destinations was outside of the scope of this work, we can examine the set of the responding intermediaries at each path outage and compare it to the set of the intermediaries that we *expect* to respond if the path failures are uncorrelated.

To shed light on the impact of the PlanetLab design, we first note that between 90% and 93% of the intermediaries in use by the vantage points in our experiment are hosted by the universities. If the connectivity failures were correlated between the universities, then we would see a noticeably smaller percentage of intermediaries hosted by universities that are valid detour nodes. In fact, across all failures on the paths originating from the university-based vantage points, we did not see significant differences between the fraction of the intermediaries hosted by the universities in use by each university-based vantage point and the fraction of their intermediaries which are located at the universities that reply as having a path to the failed destination. The average (across failures) percentage of university-hosted intermediaries in the intermediary sets of vantage points located at the universities was 92.0%, while the average percentage of university-hosted intermediaries among the intermediaries that had valid path was 91.6%. Thus, pathologies that we saw were not correlated across the paths originating at the universities.

However, similar analysis reveals geographic correlation. Across path failures at the US-based vantage points, the average percentage of intermediaries with valid paths that were located in the US was 59.8%, while the percentage of potential intermediaries that were located in the US was 71.6%. For Asian vantage points, Asian intermediaries made about 7.2% of the set of potential intermediaries, but only about 3.7% of the intermediaries with valid paths at failure. European vantage points behaved differently— the average percentage of potential intermediaries that were located in Europe was 20.8%, while the percentage of those that replied was higher, 26.7%. We do not have an explanation for this other then speculation that the European segment of the Internet is structured differently then the US and Asian segments.

Finally, we restricted a SOSR-like random-k detouring method to randomly select intermediaries from the subset that is not in the same geographical area as the vantage point. The results are reported in the right-most column of Table II under "Geographic random" selection method. While it underperforms the two deterministic methods, it outperforms the random-k significantly, especially using 5 detour nodes. Therefore, geographical path correlations on the Internet may be significant enough to be exploited by informed selection.

V. Related Work

There exists a substantial body of research on the reliability and performance of the Internet. Paxson [6] carried out one of the first Internet-scale measurement studies, notably observing the asymmetries as well as persistence in paths. Studies that followed [5], [7], [15], [16] further confirmed the path outage probability in the 1.5% to 3.3% range, as well as a heavy-tailed distribution of the outage durations. Our results are consistent with these findings. Moreover, Savage *et al.* [17] find that path selection in wide-area Internet is sub-optimal in end-to-end latency, packet loss rate, and TCP throughput spaces.

These studies motivated exploration of ways to address path selection faults and inefficiencies in the Internet. Content Distributions Networks (CDNs) [8]–[10] and clusters [18], [19] mitigate the path outages. However, their benefit is limited to only specific types of traffic, e.g. World-Wide Web. Detouring around the problematic spots in the network was also recognized as beneficial [11], leading to the development of overlay networks such as RON [1]. Much research has been devoted to improving the scalability of overlay networks [20], [21]. The basis of our work was the proposal by Gummadi *et al.* for a detouring method that did not require any path monitoring [12]. Instead, they used 4 randomly chosen nodes from a set of 67 PlanetLab machines. While they argued that choosing one detour node using information about the AS paths as seen by BGP does not result in significant increase in reliability over the four randomly chosen detour nodes, in our study we show that informed choice using recent traceroute data mapped to points-of-presence (PoPs) yields substantial path downtime reduction.

In order to conduct our study we used the traceroute data generated by the iPlane [3] project, as well as their IP prefix-to-PoP mapping. iPlane is designed to be a service that provides coarse-grained map of the Internet. Its main purpose is prediction of path properties, such as the approximate AS and PoP path between two IP prefixes of interest, latency, loss rate, bandwidth, etc. The traceroute data are one of the many inputs to their prediction engine, and are available for download on the iPlane website [14]. We used them in our project more out of convenience then necessity. While we have not utilized their predicted paths in this work due to technical difficulties, we plan on trying iPlane's and Rocketfuel's [22] path-estimation capabilities for informed detour node selection in the future.

VI. Conclusion

We presented a proposal for improving the SOSR [12] system by replacing the existing random method with an informed mechanism for selection of detour nodes and showed that it leads to a substantial improvement in reliability. We also found that biasing the SOSR random detour selection towards nodes outside of the geographic area of the source may provide an improvement in reliability at a low cost. Even though we relied on iPlane [3] as the source of data driving the detour routing decisions, we note that our system is not tied to this specific service. We believe any source of coarse-grained path information could be used for the task (though, of course, one may get a different result). While prior research suggests that most Internet paths are stable [6] over the course of 24 hours, we are curious to see whether our results would improve with fresher route data. We would also like to repeat our experiment using the paths estimated by prediction engines of iPlane and Rocketfuel [22] instead of raw traceroute data.

There are some important applications such as medical collaboration and certain financial transactions for which the current 98.5% availability of the Internet systems is not adequate. A lot of them use expensive, dedicated networks to ensure the requisite availability. While the $0.75\% - 1\%$ increase in reliability of a path that we achieve with our system may not make web-browsing or even Internet telephony performance noticeably different, the combination of a more reliable path provided by the informed detouring and increasingly reliable servers may result in these applications moving to the Internet space. Our main goal is to facilitate this by attempting to achieve the path availability improvement seen in this work in a SOSR-like system implementation.

References

[1] D. Andersen, H. Balakrishnan, F. Kaashoek, and R. Morris, "Resilient overlay networks," *SIGOPS Oper. Syst. Rev.*, vol. 35, no. 5, pp. 131–145, 2001.

[2] B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman, "PlanetLab: an overlay testbed for broad-coverage services," *SIGCOMM Comput. Commun. Rev.*, vol. 33, no. 3, pp. 3–12, 2003.

[3] H. Madhyastha, T. Isdal, M. Piatek, C. Dixon, T. Anderson, A. Krishnamurthy, and A. Venkataramani, "iPlane: an information plane for distributed services," in *Proc. of USENIX OSDI*, Nov. 2006.

[4] D. R. Kuhn, "Sources of failure in the public switched telephone network," *Computer*, vol. 30, no. 4, pp. 31–36, Apr. 1997.

[5] M. Dahlin, B. B. V. Chandra, L. Gao, and A. Nayate, "End-to-end WAN service availability," *IEEE/ACM Trans. Netw.*, vol. 11, no. 2, pp. 300–313, 2003.

[6] V. Paxson, "Measurements and analysis of end-to-end internet dynamics," Ph.D. dissertation, U.C. Berkeley, 1997.

[7] Y. Zhang, V. Paxson, and S. Shenker, "The stationarity of internet path properties: Routing, loss, and throughput," ACIRI, Tech. Rep., May 2000.

[8] B. Krishnamurthy, C. Wills, and Y. Zhang, "On the use and performance of content distribution networks," in *Proc. of ACM IMW*, Nov. 2001.

[9] J. Dilley, B. Maggs, J. Parikh, H. Prokop, R. Sitaraman, and B. Weihl, "Globally distributed content delivery," *IEEE Internet Computing*, September/October 2002.

[10] Akamai, Inc. website, http://www.akamai.com.

[11] S. Savage, T. Anderson, A. Aggarwal, D. Becker, N. Cardwell, A. Collins, E. Hoffman, J. Snell, A. Vahdat, G. Voelker, and J. Zahorjan, "Detour: informed internet routing and transport," *IEEE Micro*, vol. 19, no. 1, pp. 50–59, Jan./Feb. 1999.

[12] 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 *Proc. of USENIX OSDI*, Dec. 2004.

[13] J. Byers, J. Considine, and M. Mitzenmacher, "Geometric generalizations of the power of two choices," in *Proc. of ACM SPAA*, June 2004.

[14] iPlane website, http://iplane.cs.washington.edu.

[15] C. Labovitz, G. R. Malan, and F. Jahanian, "Internet routing instability," *IEEE/ACM Trans. Netw.*, vol. 6, no. 5, pp. 515–528, Oct. 1998.

[16] N. Feamster, D. G. Andersen, H. Balakrishnan, and M. F. Kaashoek, "Measuring the effects of internet path faults on reactive routing," in *Proc. of ACM SIGMETRICS*, Jun. 2003.

[17] S. Savage, A. Collins, E. Hoffman, J. Snell, and T. Anderson, "The end-to-end effects of internet path selection," in *Proc. of ACM SIGCOMM*, Aug. 1999.

[18] V. Cardellini, E. Casalicchio, M. Colajanni, and P. S. Yu, "The state of the art in locally distributed web-server systems," *ACM Comput. Surv.*, vol. 34, no. 2, pp. 263–311, 2002.

[19] A. Fox, S. D. Gribble, Y. Chawathe, E. A. Brewer, and P. Gauthier, "Cluster-based scalable network services," *SIGOPS Oper. Syst. Rev.*, vol. 31, no. 5, pp. 78–91, 1997.

[20] A. Nakao, L. Peterson, and A. Bavier, "Scalable routing overlay networks," *SIGOPS Oper. Syst. Rev.*, vol. 40, no. 1, pp. 49–61, 2006.

[21] C. Lumezanu, D. Levin, and N. Spring, "PeerWise Discovery and Negotiation of Faster Paths," in *Proc. of HotNets-VI*, Nov. 2007.

[22] N. Spring, R. Mahajan, D. Wetherall, and T. Anderson, "Measuring ISP topologies with rocketfuel," *IEEE/ACM Trans. Netw.*, vol. 12, no. 1, pp. 2–16, Feb. 2004.