End Point-Based Routing Strategies for Improving Internet Performance and Resilience

Aditya Akella

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Srinivasan Seshan, Chair
Bruce M. Maggs
Hui Zhang
Scott Shenker, UC Berkeley

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Copyright © 2005 by Aditya Akella

This research was sponsored by DARPA under contract F30602-99-1-0518. Additional support was provided by IBM. Views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of DARPA, NSF, IBM or the U.S. government.
**Keywords:** Internet routing, Border Gateway Protocol, performance bottlenecks, overlay networks, route control, congestion, Moore’s law, Internet topology
Abstract

Internet access speeds of large enterprises and educational institutions have improved dramatically over the past few years. However, this higher-speed connectivity is still ineffective at providing end-users with good download performance and robustness from service interruptions. This arises due to the prevalence of constrained links with little spare capacity inside Internet Service Provider (ISP) networks.

In this dissertation, we first investigate the location, latency and traffic load characteristics of network links that limit the Internet performance of well-connected end-networks. More importantly, we show how end-networks can employ a clever Internet route selection technique, called Multihoming Route Control, to avoid these performance bottlenecks and obtain much better Internet performance. Using Internet-scale measurements conducted over Akamai’s content distribution infrastructure, we show that by multihoming to three ISPs, and intelligently scheduling transfers across the ISPs, an end-network could potentially improve its Internet round-trip times (RTTs), throughputs and reliability by up as much as 30%.

We also compare the Internet performance and reliability from route control against more powerful route selection paradigms such as overlay routing. We show that the RTTs and transfer speeds from multihoming are within 5-10% of overlay routing. While multihoming cannot offer the nearly perfect resilience of overlays, we show that it can eliminate almost all failures experienced by a singly-homed end-network. We also describe the design and performance evaluation of a route control system that can be deployed by large multihomed enterprises. We show that, in practice, simple route control techniques can offer Web performance within 10% of the optimal performance from multihoming.

Finally, we investigate whether, in the future Internet, techniques such as route control or overlay routing can still provide good end-to-end performance in the face of higher access speeds and a vastly different traffic mix. We show that the structure of the Internet (i.e., a power law degree structure at the ISP level), together with the routing protocol (i.e., BGP), will turn certain keys portions of the network into persistent bottlenecks. We then consider modifications to the AS-level interconnections to guarantee good end-to-end performance in the future Internet.

We believe that the contributions in this thesis significantly advance the state-of-the-art of techniques for improving Internet performance and resilience. Further, this dissertation highlights important guidelines for the design of inter-domain routing protocols and peering architectures.
# Contents

1 Introduction 1
   1.1 Road-blocks to Efficient Internet Performance 2
       1.1.1 Non-access Bottlenecks 3
       1.1.2 Internet’s Routing Protocol: BGP 3
       1.1.3 The Problem and Our Approach 5
   1.2 Dissertation Outline 6

2 Background and Approach 9
   2.1 ISP Traffic Engineering 9
   2.2 Overlay Routing 10
   2.3 Commercial Route Control Products 12
   2.4 An Overview of Our Approach 13

3 An Empirical Evaluation of Wide-Area Internet Bottlenecks 17
   3.1 Measurement Methodology 19
       3.1.1 ISP Hierarchy 19
       3.1.2 Choosing Traffic Sources 20
       3.1.3 Choosing Probe Destinations 21
       3.1.4 Bottleneck Identification Tool: BFind 24
       3.1.5 Metrics Reported 32
   3.2 Results 32
       3.2.1 Path Properties 32
       3.2.2 Locations of Bottlenecks 35
       3.2.3 Bandwidth Characterization of Bottlenecks 36
       3.2.4 Latency Characterization of Bottlenecks 38
       3.2.5 Bottlenecks at Public Exchange Points 39
   3.3 Measurement Caveats, Summary of Observations and their Implications 40
       3.3.1 A Critique of Our Measurement Methodology 40
       3.3.2 ISPs and Provisioning 42
       3.3.3 Route Selection for Improved Internet Performance 43
4 A Measurement-Based Analysis of Multihoming 45
4.1 Naive Multihoming: RTT Performance 48
   4.1.1 Measurement Dataset 48
   4.1.2 Measurement Results 49
4.2 2-Multihoming: RTT Performance 52
   4.2.1 Measurement Dataset 52
   4.2.2 Measurement Results: 2-Multihoming 53
4.3 k-Multihoming, k ≥ 2 55
   4.3.1 Data Collection 57
   4.3.2 k-Multihoming Improvements 57
   4.3.3 Unrolling the Averages 59
   4.3.4 Impact of the Choice of ISPs 61
4.4 Resilience to Path Failures 62
   4.4.1 Active Measurements of Path Availability 62
   4.4.2 Path Availability Analysis 64
4.5 Summary of Observations and their Implications 66

5 A Comparison of Overlay Routing and Multihoming Route Control 69
5.1 Terminology 70
5.2 Latency and Throughput Performance 71
   5.2.1 Comparing RTTs and Throughputs 72
   5.2.2 1-Multihoming versus 1-Overlays 74
   5.2.3 1-Multihoming versus k-Overlays 76
   5.2.4 k-Multihoming versus 1-Overlays 76
   5.2.5 k-Multihoming versus k-Overlays 77
   5.2.6 Unrolling the Averages 79
   5.2.7 Reasons for Performance Differences 81
5.3 Resilience to Path Failures 86
   5.3.1 Active Measurements of Path Availability 86
   5.3.2 Path Availability Analysis 87
5.4 Measurement Caveats, Summary of Observations and their Implications 88

6 Practical Multihoming Route Control Strategies 93
6.1 Solution Overview 94
   6.1.1 Monitoring ISP Links 94
   6.1.2 Choosing the Best ISP 96
   6.1.3 Directing Traffic Over Selected ISPs 96
6.2 Implementation Details 97
   6.2.1 Performance Monitoring Algorithms 98
6.2.2 Switching ISPs ............................................. 101
6.2.3 NAT-based Inbound Route Control ......................... 101
6.3 Experimental Evaluation ........................................ 102
  6.3.1 Experimental Set-up ..................................... 102
  6.3.2 Experimental Results ................................... 105
6.4 Additional Design and Operational Issues ....................... 112
  6.4.1 DNS for Inbound Route Control .......................... 114
6.5 On Common Route Control Practices .......................... 115
6.6 Summary of Observations and their Implications ............... 116

7 Scaling of Congestion in the Internet .......................... 119
  7.1 Methodology .................................................. 121
    7.1.1 Problem Statement .................................... 121
    7.1.2 Simulation Set-up .................................... 123
  7.2 Analytical Results .......................................... 125
    7.2.1 Experimental Support ................................ 129
  7.3 Simulation Results .......................................... 130
    7.3.1 Shortest-Path Routing ................................ 130
    7.3.2 Policy-Based Routing ................................ 132
    7.3.3 Shortest Path Routing Variations ...................... 134
  7.4 Improving the Congestion Scaling Properties ................. 134
    7.4.1 Adding Parallel Network Links ......................... 135
  7.5 On Networking Modeling and Congestion Scaling .............. 136
  7.6 Analysis Caveats, Summary of Observation and their Implications .......................... 137

8 Conclusions and Open Problems .................................. 141
  8.1 Thesis Summary ............................................... 141
  8.2 Contributions ............................................... 141
    8.2.1 Properties of wide-area bottlenecks .................. 141
    8.2.2 Benefits of multihoming route control and comparison with overlay routing 142
    8.2.3 Route control in practice .............................. 143
    8.2.4 Congestion scaling at bottlenecks .................... 143
  8.3 Future Work ................................................ 143
    8.3.1 Longer-term Measurement Analyses ..................... 144
    8.3.2 Explaining Diminishing Returns ....................... 144
    8.3.3 Global Effects of Multihoming Route Control .......... 144
    8.3.4 New Changes to BGP ................................... 145
    8.3.5 Better Models for Congestion Scaling .................. 145
List of Figures

1.1 Evolution of home access speeds: We show when various home Internet access technologies (specifically, the corresponding access speeds) were adopted in the U.S. We also show the expected adoption dates for higher speed home access, such as 24Mbps and 100Mbps. ................................................................. 2

1.2 BGP operation: An example showing the propagation of BGP reachability and routing information across ISPs. The arrows show the propagation of routing announcements. ................................................................. 4

2.1 An end-to-end Internet flow: A typical Internet flow may traverse several ISPs end-to-end. ................................................................. 10

2.2 Overlay routing: This figure illustrates an overlay routing scenario. ............ 11

2.3 Multihoming route control: This figure illustrates a route control scenario for an end-network with three ISP connections. ................................................................. 12

3.1 ISP hierarchy: This figure illustrates the four tiers in the Internet ISP hierarchy. . 19

3.2 Locations of PlanetLab sources (a) and their connectivity properties(b): Three of our sources and seven destinations are located in Europe (shown in the inset). The size of the dots is proportional to the number of sites mapped to the same location. 21

3.3 Location and connectivity the destinations: Each destination in (a) location is identified by the PlanetLab source with minimum delay to the destination. Table (b) shows the composition of the destination set. ................................................................. 23

3.4 The operation of BFind: In either graph, queueing delay is shown on the left y-axis. The instantaneous UDP rate is shown on the right y-axis. In (a), BFind identifies hop 6 as the bottleneck. In (b), BFind identifies hop 15 as the bottleneck, although this could potentially be a false positive. ............................. 27

3.5 Topology used for BFind NS simulations: The topologies are explained in detail below. “M” stands for Mbps. The first row corresponds to location of the bottleneck link being “close”, the second corresponds to “middle” and the third to “far”. . . . . 28
3.6 **BFind interaction with competing long-lived TCP flows**: The figures plots the available bandwidth reported by BFind for the topology in Settings 1 and 2, when competing long-lived TCP flows on the bottleneck hops are constrained to at most 10Mbps.

3.7 **Relative prevalence of intra-ISP bottlenecks**: Graph (a) shows the average number of bottlenecks of each kind appearing inside ISPs, classified by path type. The graph in (b) shows the total number of links (bottleneck or not) of each kind appearing in all the paths we considered. In (a) and (b), the left bar shows the overall average number of links, while the right shows the average number of unique links. Graph (c) shows the relative fraction of intra-ISP bottlenecks links of various types (left bar) and the average path composition of all links (right bar).

3.8 **Relative prevalence of peering bottlenecks**: Graph (a) shows the average number of bottlenecks of each kind appearing between ISPs, classified by path type. The graph in (b) shows the total number of links (bottleneck or not) of each kind appearing in all the paths we considered. In (a) and (b), the left bar shows the overall average number of links, while the right shows the average number of unique links. Graph (c) shows relative fraction of peering bottlenecks of various types (left bar) and the average path composition for all links (right bar).

3.9 **Available capacity at bottleneck links**: Graph (a) corresponds to bottlenecks within ISPs. Graphs (b) and (c) show the distribution of available capacity for bottlenecks in peering links involving Tier1 ISPs, and those in peering links not involving Tier1 ISPs, respectively. We do not show the distributions for bottleneck links between tiers 2 and 4 and those between tiers 3 and 4 since they were very small in number.

3.10 **Relative prevalence of bottlenecks of various latencies**: Graph (a) shows the average number of bottlenecks of the three classes of latencies further classified into those occurring between ISPs and those occurring inside ISPs. Graph (b) shows the actual number of links (bottleneck or not) of each kind appearing in all the paths. Graph (c) shows the relative fraction of bottleneck links of various latency types (left bar) and the average path composition of all links (right bar).

3.11 **Bottlenecks in paths to exchange points**: Table (a) on the left shows the relative prevalence of bottleneck links at the exchange points. Figure (b) shows the distribution of the available capacity for bottleneck links at the exchange points.

4.1 **Multihoming**: Figure (a) shows an example of a route control set-up. Figure (b) shows a traditional multihoming set-up.
4.2 *Naive Multihoming*: Akamai servers connected to different ISPs in the same city download objects from all customer origin servers in order to serve them to clients. For this data set, turnaround times are averaged over each hour across retrievals from various origin servers.

4.3 *Naive $k$-multihoming*: Figure (a) shows the 1-multihoming performance of the ISPs in each city, with ISPs ranked according to their performance. Figure (b) shows the diminishing returns from $k$-multihoming in each city.

4.4 *Relative utilization of ISPs*: For the cities of Boston and New York, respectively, the graphs show the fraction of time the ISPs in the naive $k$-multihoming solutions at the city are utilized in the optimal schedule.

4.5 *2-Multihoming*: Akamai performance monitors in a given city are connected to different ISPs and download 10KB objects at 6-minute intervals from servers belonging to 80 content providers.

4.6 *2-multihoming evaluation*: The average benefits are shown in (a). Graph (b) shows the median, 10th and 90th percentile turnaround times for each ISP and for 2-multihoming. The relative usage of the two ISPs in the optimal schedule is shown in (c).

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

4.8 *$k$-Multihoming Benefits*: Figure (a) plots the improvement in web turnaround times from $k$-multihoming. Figure (b) plots the improvements in throughput.

4.9 *Per-destination performance*: Figures (a) and (b) plot the absolute improvements in RTT and throughput performance, respectively, from 3-multihoming relative to 1-multihoming.

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

4.11 *Time of day effects*: Figures plotting the impact of the time-of-day (Figure (a)) and day-of-week (pessimistic, Figure (b)) on RTT performance. All times are in EDT.

4.12 *Impact of sub-optimal choices*: Graph (a) shows the expected RTT performance metric from a random $k$-multihoming option. Graph (b) shows the performance of the worst $k$-multihoming option.

4.13 *Choice of ISPs*: Figures (a) and (b) show the RTT performance from various ISP selection policies for San Francisco and Los Angeles, respectively.

4.14 *End-to-end failures*: Distribution of the availability on the end-to-end paths, with and without multihoming. The ISPs in the 2- and 3-multihoming cases are the best 2 and 3 ISPs in each city based on RTT performance, respectively.
4.15 **Availability comparison**: Comparison of availability averaged across paths originating from six cities using a single ISP and using 3-multihoming. ISPs are chosen based on their round-trip time performance.

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

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

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

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

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

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

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

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

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

5.10 **“Circuitousness” of routes**: Figure plotting the propagation delay of the best indirect path (y-axis) against the best multihoming path (x-axis).
5.11 **Availability comparison:** Comparison of availability averaged across paths originating from six cities using a single ISP, 3-multihoming, 1-overlays, and 3-overlays. ISPs are chosen based on their round-trip time performance. .......... 

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

6.1 **Solution steps:** This figure illustrates the three main operations of an enterprise route control system. .......... 

6.2 **Monitoring ISP performance:** The passive measurement scheme. .......... 

6.3 **Monitoring ISP performance:** The SlidingWindow active measurement scheme. . 

6.4 **Testbed topology:** The simple test-bed, shown in (b), is used to emulate the route control scenario shown in (a). .......... 

6.5 **Web server load profile:** Average response time in ms, per KB of the request, as a function of the average client arrival rate at the server in our topology (Figure 6.4(b)). 

6.6 **Performance improvement:** The performance metric $R$ for the passive measurement scheme with EWMA parameter $\alpha = 0$ (no history employed) and sampling interval of 30s. The graph also shows the performance from the three individual ISPs. 

6.7 **Unrolling the averages:** Ratio and the difference in the response times from using just ISP 3 for all transfers relative to using the passive measurement scheme. The average client arrival rate in either case is 13.3 requests/s. .......... 

6.8 **Route control at work:** The ISPs chosen by the passive measurement-based route control scheme for destinations with different popularity levels. .......... 

6.9 **Impact of history:** The performance achieved by relying on historical samples to varying degrees. These results are for the passive measurement-based strategy with a sampling interval of 30s. .......... 

6.10 **Active vs passive measurement:** The performance of the two active measurement-based schemes, and the passive measurement scheme for a sampling interval of 120s. 

6.11 **Impact of the sampling interval:** The performance from using different sampling intervals from passive measurement-based and the FrequencyCounts active measurement-based schemes. .......... 

6.12 **DNS responsiveness:** This figure shows traffic volume over time just before and after a DNS change. The left graph (a) shows a 2-day period around the end of the event, while (b) focuses on a 2-hour period around the time of the DNS update. .. 

7.1 **Accuracy of heuristics:** The graph on the left shows the accuracy of our simple stub identification algorithm. The graph on the right shows the error in the maximum congestion due to our machine-learning based edge-classification algorithm. ..
7.2 **The model:** A pictorial view of the graph and the set $V_r$. 

7.3 (a) **Simulation support for the analytical model:** Figure (a) shows the fraction of shortest path trees that do not contain the edge $e^\alpha$. Figure (b) plots the ratio of degrees of $s_1$ and $s_2$ in a random shortest path tree to their degrees in the graph. 

7.4 **Maximum edge congestion:** Plotted as a function of $n$, in Inet-3.0 generated graphs, with $\alpha = 1.23$. The figure also plots four other functions to aid comparison – $n^{1.4}$, $n^{1.6}$, $n^{1.8}$, $n^2$. 

7.5 **Edge congestion with shortest path routing and any-2-any communication:** The figure on the left shows the maximum edge congestion. The figure on the right shows the distribution of congestion over all links, with the number of links normalized to 1 in each case. The figure on the left also plots the worst congestion for exponential graphs and preferential connectivity trees.

7.6 **Edge congestion with shortest path routing and leaf-2-leaf communication:** The figure on the left shows the maximum edge congestion. The figure on the right shows the distribution of congestion over all links (again normalized).

7.7 **Edge congestion with shortest path routing and clout model of communication:** The figure on the left shows the maximum edge congestion. The figure on the right shows the distribution of congestion over all links (again normalized).

7.8 **Policy-based routing:** Maximum Edge congestion with policy-based routing in HLSs.

7.9 **Policy-based vs shortest path routing:** Comparison of edge congestion for shortest path and policy based routing in the any-2-any model.

7.10 **Tie-breaking rules in shortest-path routing:** $\alpha = 1.23$. The figure plots the three different variations of breaking ties in shortest path routing.

7.11 **Degree vs congestion:** Edge Congestion versus the average degree of the nodes incident on the edge (any-2-any model with shortest path routing). The congestion is higher on edges with a high average degree.

7.12 **Alleviating congestion:** Maximum relative congestion for shortest path routing, any-2-any model, when parallel links are added to the graph using the sum, product and max functions.
List of Tables

2.1 **Overview of our approach:** The five key problems central to this dissertation. We also show provide citations for the preliminary conference versions of the corresponding results. ................................................................. 14

3.1 **The bandwidth-probing performance of BFind:** The table shows, for each of the six configurations of the topology in Figure 3.5(a), the output obtained from BFind and its comparison with a TCP flow on the bottleneck hop. ...................... 29

3.2 **Performance of BFind in the presence of two similar bottlenecks:** The table shows the hops identified by BFind as being the bottleneck in each of the six configurations in Figure 3.5(b), and the time taken to reach the conclusion. . . . . . . 29

3.3 **Performance of BFind in the presence of two slightly different bottlenecks:** The table shows the hops identified by BFind as being the bottleneck in each of the six configurations in Figure 3.5(b) when the bandwidth of one of the hops on the path is chosen to be slightly higher than that of the other. ......................... 30

3.4 **BFind validation results:** Statistics for the comparison between BFind, Pathload and Pipechar ......................................................... 31

3.5 **Properties of wide-area Internet Bottlenecks:** Summary of key observations regarding wide-area bottlenecks. .......................................... 41

4.1 **Rank vs overall performance** Ranks of the ISPs in the $k$-multihoming solutions at New York, $k \leq 8$, in the order in which they are added, along with the incremental performance improvement. ......................................................... 51

4.2 **Availability across router classes:** Estimated availability for routers or links classified by AS tier and location. We consider a border router as one with at least one link to another AS. ......................................................... 65

4.3 **Benefits of Multihoming Route Control:** Summary of key observations regarding multihoming. ......................................................... 67
5.1 **Throughput performance:** This table shows the 1 MB TCP transfer performance of 1-overlay routing relative to 1-multihoming (for both estimation functions). Also shown is the fraction of measurements in which 1-overlay routing selects an indirect path in each city. ................................................................. 75

5.2 **Analysis of overlay paths:** Classification of indirect paths offering > 20ms improvement in RTT performance. ................................................................. 83

5.3 **Overlay routing policy compliance:** Breakdown of the mean and 90th percentile round trip time improvement of indirect overlay routes by: (1) routes did not conform to common inter-domain policies, and (2) routes that were valid inter-domain paths but either exited ASes at different points than the direct BGP route or were different than the BGP route. ................................................................. 85

5.4 **Multihoming Route Control vs. Overlay Routing:** Summary of key comparison results. ................................................................. 89

6.1 **Characteristics of the delay traces:** Here “performance” refers to the delay along a given path. ................................................................. 105

6.2 **Analysis of performance overheads:** Here “penalty” is defined as the value of \( R - 1 \) in each case. ................................................................. 111

6.3 **Practical Multihoming Route Control:** Summary of key observations regarding route control implementation. ................................................................. 117

7.1 **Congestion Scaling in the Internet:** Summary of key observations regarding the scaling properties of the Internet. ................................................................. 138