Performance and Availability Benefits of Global Overlay Routing

Hariharan S. Rahul [*] Mangesh Kasbekar [†]

Ramesh K. Sitaraman [‡] Arthur W. Berger [§]

1 Introduction

There have been several inflection points in human history where an innovation changed every aspect of human life in a fundamental and irreversible manner. There is no doubt that we are now in the midst of a new inflection point: the Internet revolution. However, if the Internet is to realize its promise of being the next revolutionary global communication medium, we need to achieve the five grand challenges that this technology offers: perfect *availability*, high *performance*, "infinite" *scalability*, complete *security*, and, last but not the least, affordable *cost*.

As the Internet was never designed to be a mission-critical communication medium, it is perhaps not surprising that it does not provide much of what we require from it today. Therefore, significant scientific and technological innovation is required to bring the Internet's potential to fruition. Content Delivery Networks (CDNs, for short) that overlay the traditional Internet show great promise and is projected as the technology of the future for achieving these objectives.

[*] MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139 USA. rahul@csail.mit.edu.

[†] Akamai Technologies, Staines, TW18 4EP, UK. mkasbeka@akamai.com.

[‡] Department of Computer Science, University of Massachusetts, Amherst, MA 01003, USA. ramesh@cs.umass.edu.

[§] MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA and Akamai Technologies, Cambridge, MA 02142, USA. awberger@csail.mit.edu.

1.1 Architecture of CDNs Revisited

To set the context, we briefly review the evolution and architecture of commercial CDNs. Please see Chapter 1 for a more detailed overview. Before the existence of CDNs, content providers typically hosted a centralized cluster of Web and streaming servers at a data center and served content to a global audience of end users (a.k.a clients). However, this solution falls significantly short of meeting the critical requirements of availability, performance, and scalability. It suffers from both the *first-mile bottleneck* of getting content from the origin servers into the Internet, and the *middle-mile bottleneck* of transporting the content across multiple long-haul networks and peering points to the access network of the client. On the first-mile, the data center itself is a single point of failure. Any connectivity problems at the data center such as an overloaded or faulty switch can result in reduced availability or even a complete outage. On the middle mile, transporting the content over the long-haul through potentially congested peering points significantly degrades both availability and performance by increasing round-trip latencies and loss. Further, there is no protection against a flash-crowd, unless the data center is grossly over-provisioned to start with.

One can alleviate some of the shortcomings of the traditional hosting solution by *multihoming* the data center where the content is hosted [3]. This is achieved by provisioning multiple links to the data center via multiple network providers and specifying routing policies to control traffic flows on the different network links. A different but complementary approach to alleviate the problems of centralized hosting is *mirroring* the content in multiple data centers located in different networks and geographies. Both of these approaches ameliorate some of the first-mile availability concerns with centralized hosting where the failure of a single datacenter or network can bring the Website down. But, middle-mile degradations and scalability remain issues. Additionally, the operational cost and complexity are increased as multiple links and/or data centers must be actively managed. Further, network and server resources need to be overprovisioned, since a subset of the links and/or data centers must be able to handle the entire load in case of failures. As the quest for more availability and greater performance drive up the need for more multi-homed mirrors with larger server-farms, all of which mean more infrastructure costs, a CDN with a large shared distributed platform becomes attractive.

A CDN is a distributed network of servers that act as an *overlay* on top of the Internet with the goal of serving content to clients with high performance, high reliability, high scalability and low cost. A highly-simplified

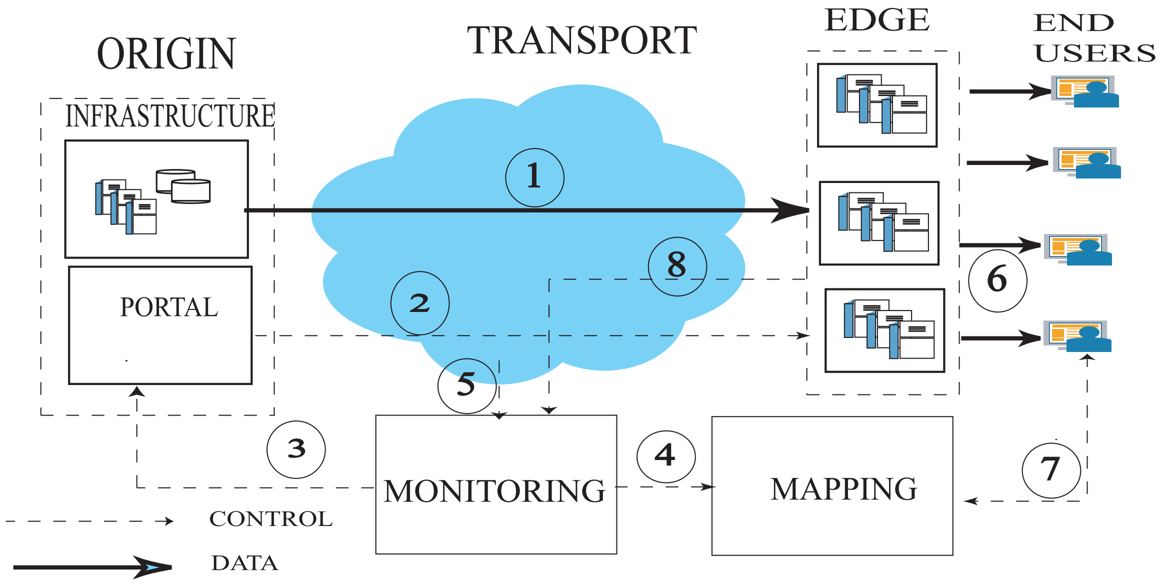

Figure 1: High-level architecture of a CDN

architectural diagram consisting of five major components is shown in Fig. 1.

Edge system. This system consist of Web, streaming, or application edge servers located close to the clients at the "edges" of the Internet. A major CDN has tens of thousands of servers situated in thousands of networks (ISPs) located in all key geographies around the world. The edge system downloads content from the origin system (Arrow 1 in Fig. 1), caches it when relevant, and serves it out to the clients. A more sophisticated system may also perform application processing to dynamically construct the content at the edge before delivering it to the client.

Monitoring system. This system monitors in real-time both the "Internet weather" and the health of all the components of the CDN, including the edge servers. Input (5) in Fig. 1 from the Internet cloud could consist of slow-changing information such as BGP feeds from tens of thousands of networks, and fast-changing performance information collected through traceroutes and "pings" between hundreds of thousands of points in the Internet. Input (8) consists of detailed information about edge servers, routers, and other system components, including their liveness, load, and resource usage.

Mapping system. The job of the mapping system is to direct clients to their respective "optimal" edge servers to download the requested content (Arrow 6). The common mechanism that mapping uses to direct clients to their respective target edge servers is the Domain Name System (DNS, Arrow 7). Typically, a content provider's domain www.cp.com is aliased (i.e. CNAME'd) to a domain hosted by the CDN, such as www.cp.com.cdn.net. A name lookup by a client's nameserver of the latter domain results in the target server's ip being returned [10]. Mapping must ensure that it "maps" each client request to an "optimum" target server that possesses the following properties: (a) the target server is live and is likely to have the requested content and is capable of serving it; (b) the target server is not overloaded, where load is measured in terms of CPU, memory, disk and network utilization; (c) the target server has good network connectivity to the client, example, little or no packet loss and small round-trip latencies. To make its decisions, mapping takes as input both the Internet weather and the condition of the edge servers from the monitoring system (Input 4), and an estimate of traffic generated by each nameserver on the Internet and performs a complex optimization to produce an assignment.

Transport system. This system is responsible for transporting data over the long-haul across the Internet. The types of content transported by the system is varied and have different quality-of-service requirements, which makes the design of this system very challenging. For instance, transporting live streaming content from the origin (i.e. encoders) to the edge servers has a different set of requirements, as compared to transporting dynamic Web content from origin to the edge. The challenge of course is designing a small and maintainable set of general-purpose mechanisms and abstractions that can satisfy the diverse requirements.

Origin system. This system originates the content that is served out to a global audience of the clients, and as such a large CDN could have tens of thousands of origin systems (one or more per content provider) that interact with the rest of the CDN. The origin Web infrastructure may include application, database, and Web servers. The origin infrastructure for streaming media could include large fault-tolerant replicated storage servers for storing on-demand (i.e. pre-recorded) content or equipment for video capture and encoding for live content. The origin infrastructure is usually (but not always) operated by the content provider, typically out of a single data center that is in a some cases multihomed and/or mirrored. The origin system also includes the portal operated by the CDN that is the "command center" for the content provider to provision and control their content (Arrows 2 and 3).

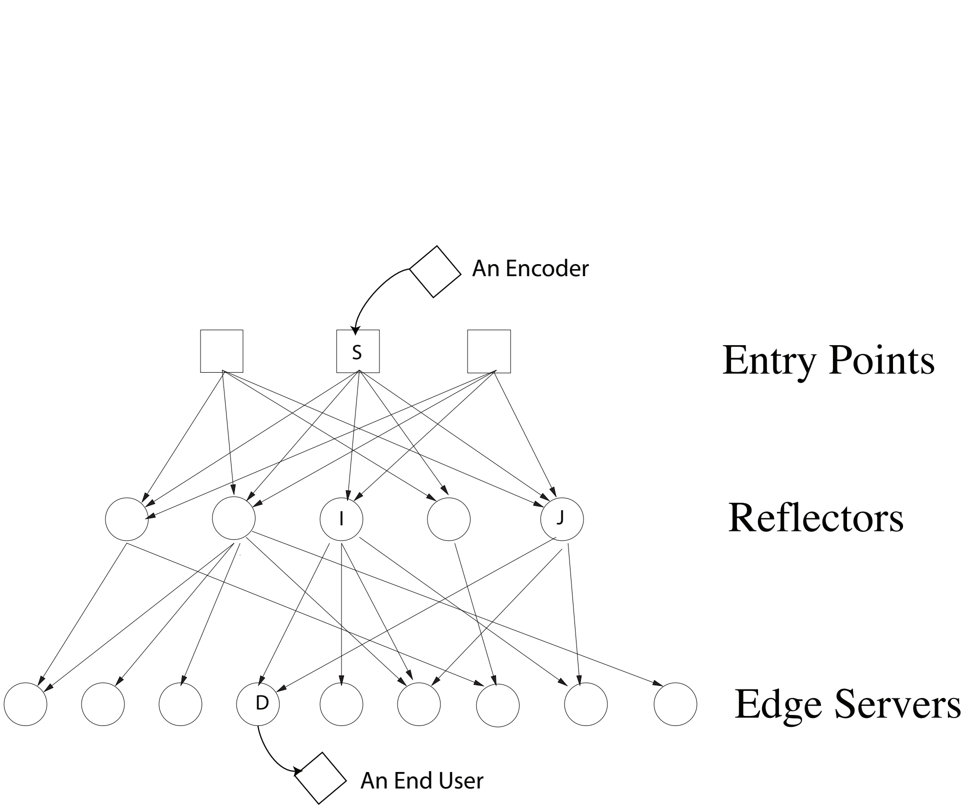

Figure 2: A transport system for live streaming

1.2 Transport Systems

In this section, we review different types of transport systems and the optimizations that they perform to enhance performance. A transport system is distinguished by the end-to-end requirements of the transported content. We review some of the optimizations performed by transport systems.

1.2.1 Live Streaming

A transport system for live streaming transmits live media content from the source of the stream (encoder) to the end user, so as to optimize the end user's experience of the stream (See Fig. 2). An *encoder* encodes the live event and sends out a sequence of encoded data packets for the duration of the live event. This data stream is first sent from the encoder to a cluster of servers called the *entry point*. It is important that the entry point can be reached from the encoder with low network latency and little or no loss. The connectivity between the encoder and its entry point is constantly monitored, and if the connectivity degrades or if the entry point fails for any other reason, the transport system automatically diverts the stream to a different entry point that is functioning well. From the entry point, the stream is sent to one or more server clusters called *reflectors*. Each reflector, in turn, sends the data stream to one or more edge server clusters. Finally, each end user obtains the live stream from a nearby edge server using the mapping system.

The goal of the transport system is to transmit live streams in a manner that stream quality is enhanced and distortions are minimized. Distortions experienced by end users include large delays before the stream starts up, information loss leading to degraded audio and video, and freezes during playback. Each stream is sent through one or more paths adaptively by using the reflectors as intermediate nodes between the entry point and the edge server. As an example, the stream entering entry point S can be duplicated across one path through reflector I and an additional path through reflector J to reach edge server D (see Fig. 2). If a data packet is lost on one path, the packet may be recovered at the edge if its duplicate is received through the other path. A more sophisticated technique would be to use a coding scheme to encode the data packets, and send the encoded stream across multiple paths. Even if some packets are lost in transit, they may be recovered at the edge servers using a decoding process.

Another example of an optimization is *prebursting*, where the initial portion of the stream is transported to the end user at a rate higher than the encoded bit rate, so as to fill the buffer of the end user's media player quickly. This allows the media player to start the stream up quicker and also decreases the likelihood of a freeze in the middle of a playback. For more discussion of the algorithmic and architectural issues in the design of streaming transport systems, see [6] and [12] respectively.

1.2.2 Web and Online Applications

A transport system for the Web transports dynamically-generated content between the origin and the edge. Such content includes both dynamic Web pages downloaded by the end user and user-generated content that is uploaded to a Website. A goal of such a transport system is to optimize the response times of Web transactions performed by the end users. As with streaming, the transport system may use one more intermediate nodes to efficiently transmit information from the origin to edge. Also as with streaming, the transport system performs several application-specific optimizations. For instance, a transport system for accelerating dynamic Web content may prefetch the embedded content on a Web page from the origin to the edge, so as to "hide" the communication latency between the origin and the edge.

A transport system for ip-based applications is focused on accelerating specific (non-http) application technologies such as Virtual Private Networks

(VPNs) and Voice-over-IP (VOIP). The architectural issues in such systems are qualitatively different from that of the Web due to the highly-interactive real-time nature of the end user experience.

1.2.3 Overlay Routing Schemes

A transport system uses a number of *application-specific* enhancements to meet the end-to-end requirements. For instance, as noted, transport systems use coding for loss recovery, prebursting for fast stream startup, and prefetching for fast downloads [6, 12]. These types of application-specific enhancements play a significant part of the overall performance benefit offered by the transport system, but are not the focus of the empirical study presented in this chapter. However, a fundamental benefit of *all* transport system is finding a "better path" through the Internet from the point where the content originates (origin, encoder, etc) to the point where the content is served to the end user (edge). This purely network-level benefit is achieved through an *overlay routing scheme* that is implemented as a part of the transport system.

A generic overlay routing scheme computes one or more *overlay paths* from each source node S (typically the origin) to each destination node D (typically the edge server) such that the overlay path(s) have high availability and low latency. The overlay routing scheme typically computes overlay paths for millions of source-destination pairs using Internet measurement data. Often, the BGP-determined Internet path from a source S to a destination D, also called the *direct path*, is not the "best path" between those two nodes. This should not be surprising as the Internet protocols that select the route are largely policy-based rather than performance-based. It could well be that an *indirect path*[1] that goes from S to an intermediate node I (typically another server cluster belonging to the CDN) and then goes from I to D is faster and/or more available! An overlay routing scheme exploits this phenomenon to choose the best overlay path (direct or indirect) to route the content, thereby enhancing the end-to-end availability and performance experienced by the end user. The benefits of a global overlay routing schemes is our focus for the rest of this chapter.

1.3 Our Contributions

We present an empirical evaluation of the performance and availability benefits of global overlay routing. There has been much recent work [4, 11, 22]

[1] An indirect path may have more than one intermediate node if necessary.

on improving the performance and availability of the Internet using overlay routing, but they have one of the following limitations:

- Prior work was performed on a platform hosted largely on Internet2[2], whose capacity and usage patterns, as well as policies and goals, differ significantly from the commercial Internet.
- Overlays used in prior work have a footprint primarily in North America. However, it is well known that network interconnectivity and relationships in Europe and Asia are different than the continental United States.

In this chapter, we present the results of the first empirical study of the performance and availability benefits of routing overlays on the commercial Internet. We use a global subset of the Akamai CDN for data collection. Specifically, we collect measurements from 1100 locations distributed across many different kinds of ISPs in 77 countries, 630 cities, and 6 continents. We address the problem of picking optimum overlay paths between the edge servers situated near end users and origin servers situated in the core of the Internet. We investigate both performance characterized by round trip latency as well as path availability. Applications such as large file downloads whose performance is more accurately characterized by throughput are not addressed in this study.

The key contributions of this chapter are the following:

- It is the first evaluation of an overlay that utilizes data from the commercial Internet. Our study provides useful cross validation for the currently deployed testbeds such as PlanetLab [18] and RON [22], and indicates that, while these deployments provide qualitatively similar data for the commercial Internet in North America, they do not capture the global diversity of network topology, especially in Asia.
- We show that randomly picking a small number of redundant paths (3 for Europe and North America, and 5 for Asia) achieves availability gains that approach the optimal. Additionally, we demonstrate that for reasonable probing intervals (say, 10 minutes) and redundancy (2 paths), over 90% of the source-destination pairs outside Asia have latency improvements within 10% of the ideal, whereas paths that originate or end in Asia require 3 paths to reach the same levels of performance.

[2] Internet2 is an advanced networking consortium consisting of several major research and educational institutions in the US. Internet2 operates an IP network that can be used for research purposes.

- We provide strong evidence that overlay choices have a surprisingly high level of persistence over long periods of time (several hours), indicating that relatively infrequent network probing and measurements can provide optimal performance for almost all source-destination pairs.

1.4 Roadmap

The rest of the chapter is organized as follows. Section 2 presents an overview of related work, and outlines the context of our present study. Section 3 describes our testbed and how the measurement data is collected. Sections 4 and 5 provide detailed metrics on the ideal performance and availability gains that can be achieved by overlays in a global context. Section 6 addresses issues in real overlay design, and explores structural and temporal properties of practical overlays for performance and availability. In Sections 7 and 8, we provide directions for further research and a vision for the future.

2 Related work

There have been many measurement studies of Internet performance and availability, for example, the work at the Cooperative Association for Internet Data Analysis (CAIDA) [7], and the National Internet Measurement Infrastructure (NIMI) [16, 17]. Examples of routing overlay networks built in academia include the Resilient Overlay Networks project at MIT [22] and the Detour project at U. Washington [11]. Commercial delivery services offered by Akamai Technologies [1] incorporate overlay routing for live streaming, dynamic Web content, and application acceleration.

Andersen *et al.* [5] present the implementation and performance analysis of a routing overlay called Resilient Overlay Networks (RON). They found that their overlay improved latency 51% of the time, which is comparable to the 63% we obtain for paths inside North America. Akella *et al.* [2] investigate how well a simpler route-control multi-homing solution compares with an overlay routing solution. Although the focus of that study is different from our current work, it includes results for a default case of a single-homed site, and the authors find that overlay routing improves performance as measured by round-trip latency by 25% on average. The experiment was run using 68 nodes located in 17 cities in the U.S., and can be compared with the 110 node, intra-North-America case in our study, where we find that the overall latency improvement is approximately 21%. However, we show that the improvement varies significantly for other continents. Savage *et al.* [23]

used data sets of 20 to 40 nodes and found that for roughly 10% of the source-destination pairs, the best overlay path has 50% lower latency than the direct path. We obtain the comparable value of 9% of source-destination pairs for the case of intra-North America nodes, though again significantly disparate results for other continent pairs. In parallel with our evaluation, Gummadi *et al.* [13] implemented random one-hop source routing on PlanetLab and showed that using up to 4 randomly chosen intermediaries improves the reliability of Internet paths.

3 Experimental Setup

We describe the experimental setup for collecting data that can be used to optimize Internet paths between edge networks where end users are located and enterprise origin servers. End users are normally located in small lower-tier networks, while enterprise origin servers are usually hosted in tier-one networks. We consider routing overlays comprised of nodes deployed in large tier-one networks that function as intermediate nodes in an indirect path from the source (enterprise origin server) to the destination (edge server).

3.1 Measurement Platform

The servers of the Akamai CDN are deployed in clusters in several thousand geographic and network locations. A large set of these clusters is located near the edge of the Internet, i.e. close to the end users in non-tier-one providers. A smaller set exists near the core ISPs directly located in tier-one providers, i.e. in locations that are suitable for enterprise origin servers. We chose a subset of 1100 clusters from the whole CDN for this experiment, based on geographic and network location diversity, security, and other considerations. These clusters span 6 continents, 77 countries, and 630 cities. Machines in one cluster get their connectivity from a single provider. Approximately 15% of these clusters are located at the core, and the rest are at the edge. The intermediate nodes of the overlay (used for the indirect paths) are limited to the core set. Table 1 shows the geographic distribution of the selected nodes. All the data collection for this work was done in complete isolation from the CDN's usual data collection activity.

3.2 Data Collection for Performance and Availability

Each of the 1100 clusters ran a task that sent ICMP echo requests (pings) of size 64 bytes every 2 minutes to each node in the core set (this keeps the

Continent (Mnemonic)	Edge set	Core set
Africa (AF)	6	0
Asia (AS)	124	11
Central America (CA)	13	0
Europe (EU)	154	30
North America (NA)	624	110
Oceania (OC)	33	0
South America (SA)	38	0

Table 1: Geographic distribution of the platform

rate of requests at a core node to less than 10 per second). Each task lasted for 1.5 hours. If a packet was lost, specifically if no response is received within 10 seconds, then a special value was reported as the round-trip latency. Three tasks were run every day across all clusters, coinciding with peak traffic hours in East Asia, Europe, and the east coast of North America. These tasks ran for a total of 4 weeks starting 18 October, 2004. Thus, in this experiment, each path was probed 3,780 times, and the total number of probes was about 652 million. A small number of nodes in the core set became unavailable for extended periods of time due to maintenance or infrastructure changes. A filtering step was applied to the data to purge all the data for these nodes. A modified version of the standard all-pairs shortest path algorithm [9] was executed on the data set to determine the shortest paths with one, two, and three intermediate nodes from the core set. We obtained an archive of 7-tuples `<timestamp, source-id, destination-id, direct RTT, one-hop shortest RTT, two-hop shortest RTT, three-hop shortest RTT>`. The archive was split into broad categories based on source and destination continents.

We consider a path to be unavailable if three or more consecutive pings are lost. Akella *et al.* [2] use the same definition, where the pings were sent at one minute intervals. The alternative scenario that three consecutive pings are each lost due to random congestion occurs with a probability of order 10^{-6}, assuming independent losses in two minute epochs with a probability of order 1%. We consider the unavailability period to start when the first lost ping was sent, and to end when the last of the consecutively lost pings was sent. This is likely a conservative estimate of the length of the period, and implies that we only draw conclusions about Internet path failures of duration longer than 6 minutes.

We filtered out all measurements originating from edge nodes in China for our availability analysis. Their failure characteristics are remarkably different from all other Internet paths as a consequence of firewall policies applied by the Chinese government.

3.3 Evaluation

We aggregate our results based on the continents of the source and destination nodes, motivated by the fact that enterprise Websites tend to specify their audience of interest in terms of their continent. The categories are denoted by obvious mnemonics such as AS-NA (indicated in Table 1), denoting that the edge servers are in Asia and origin servers are in North America.

4 Performance Benefits of Overlay Routing

In this section, we evaluate the performance benefits of overlay routing in the ideal situation where all possible indirect paths are considered for each source-destination pair, and the optimal indirect path is chosen in real time. Recall that our metric of performance is latency which is the round-trip time (abbreviated to RTT) from source to destination.

We compare the direct and the fastest indirect path for each source-destination pair and present the results in Table 2. We divide the data set into buckets based on its category and the percentage improvement in the latency of the fastest indirect path as compared to the direct path. Table 2 shows the percentage of source-destination pairs that fell in each of the buckets. The rows of the table sum to 100%. As an explanatory example for Table 2, consider the AS-AS row. The "$< -10\%$" bucket shows the cases where the best indirect paths are at least 10% slower than the direct path. 15.5% of the AS-AS paths fell in this bucket. The "$\pm10\%$" bucket represents the cases where the best indirect path and the direct path are comparable, in the sense that their latencies are within 10% of each other. 24.7% of the paths in the AS-AS category fell in this bucket. Out of the remaining direct paths, 23.4% saw a marginal (10-30%) improvement, 13.2% of the paths saw significant (30-50%) improvements, and 23.2% of the paths saw large latency reductions of a factor of two or better from the indirect paths found by the overlay.

Overall, about 4% to 35% of all source-destination pairs see improvements of over 30% latency, depending on the category. Additionally, high numbers of source-destination pairs see over 50% improvement for the AS-AS and EU-EU categories, which indicates the presence of many cases of pathological routing between ISPs in these continents. A nontrivial number

Category	$< -10\%$ (Slower)	$\pm 10\%$ (Comparable)	$10 - 30\%$ (Marginal)	$30 - 50\%$ (Significant)	$> 50\%$ (Large)
AF-AS	4.0	44.5	44.2	5.7	1.6
AF-EU	0.6	69.3	18.1	9.7	2.3
AF-NA	0.0	74.2	21.6	3.5	0.6
AS-AS	15.5	24.7	23.4	13.2	23.2
AS-EU	0.9	33.9	45.5	12.5	7.2
AS-NA	0.1	43.2	42.4	7.6	6.7
CA-AS	0.0	40.5	53.5	4.6	1.4
CA-EU	1.4	53.2	42.3	2.5	0.7
CA-NA	1.7	44.1	41.3	11.2	1.8
EU-AS	0.6	24.5	63.8	7.8	3.2
EU-EU	10.5	36.4	30.5	12.6	10.0
EU-NA	0.0	50.6	45.1	3.3	0.9
NA-AS	0.0	34.0	57.9	5.4	2.6
NA-EU	0.1	43.1	51.1	4.4	1.4
NA-NA	2.4	34.7	39.0	15.0	9.0
OC-AS	6.1	38.9	18.9	22.9	13.2
OC-EU	0.0	60.4	35.1	3.9	0.7
OC-NA	0.0	66.7	25.6	6.3	1.4
SA-AS	0.1	43.1	47.9	5.5	3.4
SA-EU	0.4	66.1	28.9	2.3	2.2
SA-NA	0.9	55.1	35.1	5.7	3.3

Table 2: Histogram of latency reduction percentages

of AS-AS paths are routed through peering locations in California, for example, the path between Gigamedia, Taipei and China Telecom, Shanghai. All the traceroutes in our snapshot that originated at Gigamedia, Taipei and ended at other locations in Asia went via California, except the path to China Telecom, Shanghai, which went directly from Taipei to Shanghai. The Taipei-Shanghai path thus sees little or no improvement with an overlay, since all the alternatives are very convoluted. At the same time, all the paths that originate in Gigamedia, Taipei and end in other locations in Asia see *large* improvements, since their direct routes are very convoluted, but there exists a path via China Telecom, Shanghai, which is more than 50% faster.

Category	50th percentile			90th percentile		
	Direct (ms)	Fastest (ms)	Reduction (%)	Direct (ms)	Fastest (ms)	Reduction (%)
AF-AS	350	290	17	740	700	5
AF-EU	150	120	20	620	620	0
AF-NA	200	180	10	560	550	2
AS-AS	230	110	52	590	350	41
AS-EU	320	260	19	500	360	28
AS-NA	230	200	13	470	280	40
CA-AS	230	200	13	300	250	17
CA-EU	160	140	12	200	170	15
CA-NA	90	70	22	130	100	23
EU-AS	300	260	13	390	300	23
EU-EU	30	30	0	80	60	25
EU-NA	130	120	8	190	160	16
NA-AS	190	160	16	260	210	19
NA-EU	130	110	15	180	150	17
NA-NA	50	40	20	90	70	22
OC-AS	200	140	30	340	220	35
OC-EU	330	300	9	400	330	17
OC-NA	220	200	9	280	230	18
SA-AS	320	280	12	470	340	28
SA-EU	230	210	9	290	250	14
SA-NA	160	150	6	240	190	21

Table 3: Latency reduction for typical and poorly-connected source-destination pairs

4.1 Source-Destination Pairs with Poor Connectivity

Enterprises are particularly interested in enhancing the worst-case performance of their Website, by speeding up the clients who see the worst performance. Therefore, the benefits provided by overlay routing in minimizing the worst path latencies in each category are especially interesting. We compare the latency reduction enjoyed by a “typical” source-destination pair in a given category with that of a “poorly connected” source-destination pair in the same category. We bucketed the data set for each category into 10 millisecond buckets based on the latency of the direct path. We then looked at the 50th percentile bucket (“typical” source-destination pairs) and the 90th

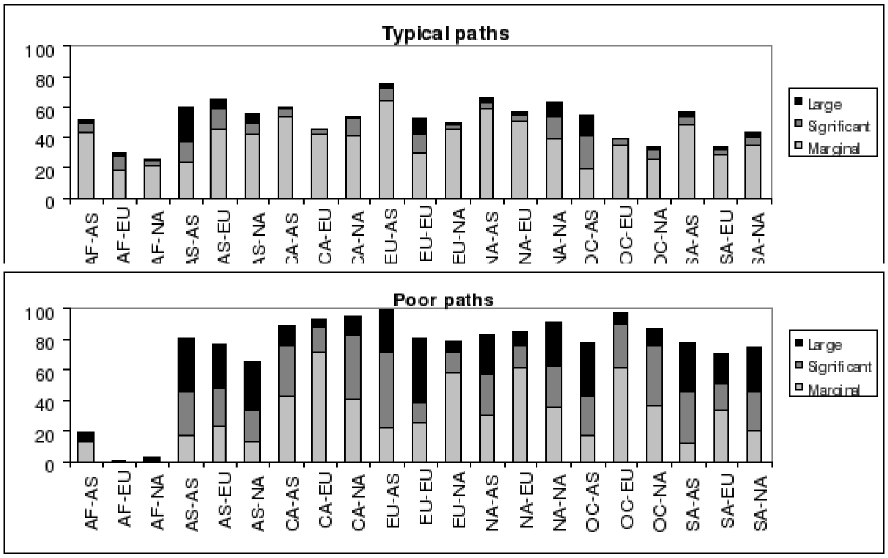

Figure 3: Latency reduction for all and poorly-connected source-destination pairs

percentile bucket ("poorly-connected" source-destination pairs). For each of these buckets, we determined the average improvements provided by the fastest indirect path over the direct path. Table 3 shows the comparison of the benefits seen by the typical and the poorly-connected source-destination pairs in each category. For the typical source-destination pairs, the latency reduction exceeds 20% only for AS-AS, OC-AS and CA-NA out of the 21 categories. Comparatively, the poorly-connected source-destination pairs see a benefit over 20% for half of the categories. The important categories of AS-AS, AS-NA, and EU-EU show significant improvements for the poor source-destination pairs, while, in contrast for paths originating from Africa the latency for 90th percentile bucket is both high and not helped with the overlay. For the AS-AS category, both the typical and poor source-destination pairs see significant improvement via the overlay, but the improvement are even greater for the typical paths. However, in general we can conclude that poorly-connected source-destination pairs benefit more from overlay routing, compared to a typical source-destination pair.

Next, we provide a more in-depth evaluation of what fraction of the poorly-connected source-destination pairs derive marginal, significant, or a large benefit from overlay routing. We bucket all the source-destination pairs in a given category whose direct path latency ever exceeded the 90th percentile latency of that category as shown in Table 3 to derive the histogram of the latency reduction for poorly-connected source-destination pairs. This histogram of the latency reduction for poorly-connected source-destination pairs is shown along side the same values for all source-destination pairs in that category in Fig. 3. (Note that the data charted in Fig. 3 for all source-destination pairs was presented in the last three columns of Table 2). Poorly-connected source-destination pairs see at least marginal benefits in over 80% of the samples, while 67% of the samples see significant or large benefits. Some categories do deviate from this observation in the figure. For example, even poorly-connected source-destination pairs with destinations in Africa do not derive much help from an overlay.

5 Availability Gains of Overlays

In this section, we evaluate the availability benefits of overlay routing in the ideal situation, where all possible indirect paths are considered for each source-destination pair, and when possible an indirect path that is available is chosen in real time to mitigate failures.

We study how often the direct path from each source-destination pair fails, and during these failures what percentage of times at least one indirect path was functional. This provides a best-case estimate of the availability gains that overlay routing can provide. Fig. 4 shows the percentage samples where the direct path between the source and destination failed for each category. The failure percentage of the direct paths ranges from 0.03% to 0.83%. Asia has the poorest availability: nine of the ten categories with the largest failure percent have an endpoint in Asia. In the presence of overlay routing, the failure percent goes down by 0.3-0.5% for most categories, indicating that the indirect paths help mask failures of the direct path. In fact, the high-failure categories involving Asia show dramatic availability improvements.

5.1 Source-Destination Pairs with Poor Connectivity

As with Section 4.1, we study how overlay routing benefits source-destination pairs with direct paths that exhibit the most failures. Again, this is of great interest to enterprises that are typically interested in using CDNs to enhance the availability of their least available end users and clients. It is commonly understood that a small number of paths contribute to a large number of path failures on the Internet. As evaluated in [15], 3% of Internet paths give rise to 30% of failures. We identified a similar pattern in our data as shown in Table 4. We see that about 3% of the direct paths caused 30% of the failures, and that 10% of the direct paths gave rise to 50% of the failures.

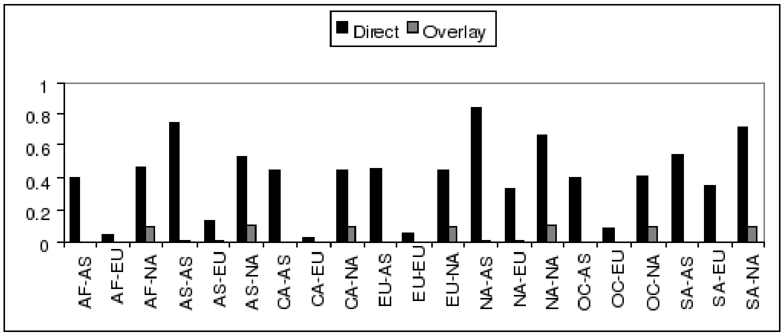

Figure 4: Reduction in failure percentages with overlay routing

Category	% paths with 30% failures	Failure % no overlay	Failure % overlay
AF-AS	4.5	25.8	0
AF-EU	1.7	8.8	0
AF-NA	0.6	36.2	0
AS-AS	2.7	31.4	0
AS-EU	1.5	9.8	0
AS-NA	0.4	30	0
CA-AS	3.5	28.2	0
CA-EU	1.6	10.9	0
CA-NA	0.5	30.3	0
EU-AS	3	30.1	0
EU-EU	0.9	10.8	0
EU-NA	0.4	30.1	0
NA-AS	2.7	32.3	0
NA-EU	0.4	13.2	0
NA-NA	0.2	40.2	0
OC-AS	3.1	30.8	0
OC-EU	1.4	10.7	0
OC-NA	0.4	29.3	0
SA-AS	3.3	28.8	0
SA-EU	2.2	9.5	0
SA-NA	0.8	23	0

Table 4: Availability statistics for poor paths

We identified the least-available source-destination pairs in each category that cumulatively gave rise to 30% of the failures, and re-ran the availability analysis for only these source-destination pairs. The results are shown in Table 4. A failure rate higher than 20% for direct paths for a source-destination pair is indicative of some specific chronic trouble, rather than random, transient failures or short-lived congestion. Almost all these source-destination pairs with a chronic availability problem saw perfect availability with overlay routing! Enhancing the availability of the least available origin-destination pairs is a key benefit of overlay routing.

6 Achieving the Benefits in a Practical Design

The analysis presented in Sections 4 and 5 characterizes an ideal case where network measurements are used in the computation of indirect paths in real-time. In addition, we assumed that an unlimited number of indirect paths can be probed and utilized as indirect routes. Therefore, this analysis is a best-case estimate on the performance and availability gains that can be expected from overlay routing. However, in a practical system, measurements made at a given time t is used for constructing overlay paths that are utilized by the transport system till some time $t + \tau$ into future. And, only a small number of indirect paths can be constructed and used at any given time for a given source-destination pair (call the number of paths κ). This section incorporates these practical considerations into the analysis and evaluates its impact on the results. As κ increases and τ decreases, the cost of constructing the overlay paths goes up but one would expect the quality of constructed overlay paths to increase and approach the best-case routes constructed in Sections 4 and 5.

First, we evaluate a simple multi-path memoryless overlay routing scheme that *randomly* selects a subset of κ paths based purely on static information and uses it to route content. It is natural to expect that this overlay will likely be inferior to the ideal, but our goal is to develop a straw man to validate the importance of intelligence and adaptiveness in overlay path selection. Surprisingly, we found that random selection is successful in providing near optimal availability for $\kappa = 3$, substantiating the fact that the Internet offers very good path diversity, and generally has low rates of failure. The policy, however, fails in improving performance, suggesting that careful path selection is very important in building overlays for performance gains. Such performance-optimizing overlay routing schemes are the focus of the rest of the section.

6.1 Stability of Optimal Paths

To the extent that a performance-optimizing overlay routing scheme selects a subset of paths to use, it will deviate from optimality as a result of variations in path latencies over time that cause a reordering of the best paths. Source-destination pairs tend to fall into two categories:

1. The best paths from the source to the destination are quite persistent, and do not change, regardless of variations in the latencies of all paths between them.

2. Latency variations of the paths over time cause a significant reordering of the best paths between source and destination, which in turn causes changes in the optimal paths.

Source-destination pairs in the first category do not require a very dynamic overlay design for selecting indirect paths for performance improvement. For example, consider the path from Pacific Internet, Singapore to AboveNet, London. The direct path, which hops from Singapore through Tokyo, San Francisco, Dallas, and Washington D.C. to London takes approximately 340 msec. However, there exists an indirect path through an intermediate node in the ISP Energis Communications in London. The path between Pacific Internet, Singapore and Energis, London is one hop long (possibly a satellite link), and has a latency of 196 ms. The subsequent traversal from Energis, London to AboveNet, London takes just 2 ms. The indirect path is therefore faster than the direct path by over 140 ms, or 41.2%. While the latencies vary, the ordering of the paths seldom change.

For source-destination pairs in the second category, latency variations are more important. We systematically examine the extent of the latency variation across paths by computing a statistic called *churn* that measures the extent to which sets of best κ paths at two different time instants vary. Formally, for a given pair of nodes,

$$Churn_t(\kappa, \tau) \overset{\Delta}{=} |S(\kappa, t) - S(\kappa, t + \tau)|/\kappa,$$

where $S(\kappa, t)$ is the set of the κ best performing paths between those nodes at time t. $Churn(\kappa, \tau)$ for a node pair is then computed as an average of $Churn_t(\kappa, \tau)$ over all valid values of t. $Churn(\kappa, \tau)$ is a number between 0 and 1, that is 0 for paths with a persistent set of best paths, and tend to be closer to 1 for paths with a fast changing set of best paths. We found that the majority of source-destination pairs have values of $Churn(\kappa, \tau)$ larger

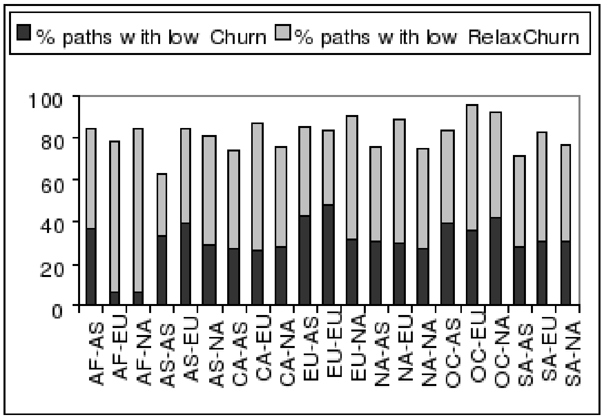

Figure 5: Percentage of source-destination pairs with low $Churn$ and $RelaxChurn$ for $\tau = 2$ minutes and $\kappa = 1$

than 10%, even when selecting up to $\kappa = 5$ best performing paths and using this prediction for only $\tau = 2$ minutes into the future.

To examine path churn more closely, one can define a relaxed measure called $RelaxChurn(\kappa, \tau)$ that counts only paths $\pi \in S(\kappa, t) - S(\kappa, t + \tau)$ whose latency at $t + \tau$ is higher than 110% of the latency of the path with the worst latency in $S(\kappa, t + \tau)$, i.e. keeping path π would worsen the performance at time $t + \tau$ by more than 10%. Interestingly, $RelaxChurn(\kappa, \tau)$ is less than 10% on average for over 80% of source-destination pairs in most categories. This indicates that a path selection algorithm that makes predictions into the future based on current measurements, can achieve performance close to the ideal.

Fig. 5 shows the percentage of source-destination pairs that have $Churn(\kappa, \tau)$ and $RelaxChurn(\kappa, \tau)$ of less than 10% for $\kappa = 1$ and $\tau = 2$ minutes. Note that paths with both the end points in Asia do have a higher value of $RelaxChurn$ than $Churn$, but still only 63% AS-AS source-destination pairs have low-churn paths. Thus, potentially higher performance benefits for AS-AS paths are likely only obtainable at a higher cost in terms of network measurement.

6.2 Performance Gains of a Predictive Overlay

The analysis in Section 6.1 examined stability using purely structural properties. In this section, we compare the performance of overlay routing with parameters κ and τ with the performance of the ideal case where the optimal path is always chosen. Note that this measure holds overlays to a higher

Category	Percentage of paths			
	$\kappa = 1$ $\tau = 2$	$\kappa = 1$ $\tau = 10$	$\kappa = 2$ $\tau = 2$	$\kappa = 3$ $\tau = 2$
AS-AS	62.4	59.5	84.6	89.4
AS-EU	76.2	74.1	92.2	94.5
AS-NA	74.8	71.6	94.0	96.0
EU-AS	74.4	72.3	88.4	92.8
EU-EU	80.1	78.1	91.6	93.1
EU-NA	83.0	82.2	94.7	96.2
NA-AS	68.1	66.2	88.8	93.7
NA-EU	82.3	81.3	95.4	97.2
NA-NA	71.6	69.6	92.0	95.0

Table 5: Percentage of paths within 10% of the optimal latency

standard, as the optimal path at a given time is at least as fast as the direct path.

A natural case to examine in some detail would be $\kappa = 1$. This corresponds to just using the best path choice in future iterations. Table 5 in the second and third columns shows our results for $\tau = 2$ and 10 minutes. As an explanatory example, consider the NA-NA category. The table shows that when using $\tau = 2$ minutes, 71.6% of the paths came within 10% of the optimal latency for that observation. Even when using stale data, with $\tau = 10$ minutes, 69.6% of the paths managed to achieve the same result. Paths originating in Asia again show a greater deviation from optimality than paths originating in Europe, whereas paths originating in North America span the full range of deviations.

Given that the performance gains with $\kappa = 1$ do not seem adequate everywhere, we then explored higher values of κ. As an explanatory example, consider the category NA-EU. The table shows that 82.3% of the paths came within 10% of the optimal when choosing $\kappa = 1$. Increasing κ to 2 enabled approximately 13.1% more paths to achieve the same result. Increasing κ to 3 provides only a marginal benefit for the remaining paths, and only 1.8% more paths achieved the result with this value of κ. From Table 5, we immediately see that choosing $\kappa = 2$ provides disproportionately high gains over choosing $\kappa = 1$, and the marginal benefit of choosing $\kappa = 3$ is much lower. In fact, apart from paths with their destination in Asia, over 90% of all source-destination pairs are within 10% of the ideal performance when selecting $\kappa = 2$, and this fact remains true even with increasing τ.

The results also suggest that an overlay routing scheme where either $\kappa = 1$ or 2 paths are used would work well. For example, 95.4% of all NA-EU source-destination pairs are within 10% of optimal for overlays with $\kappa = 2$. Combining this with the fact that 82.3% of these pairs require only one choice to come within the same limits, it is conceivable that an overlay routing scheme could potentially use two paths only for the excess 13.1% of pairs, for an average overhead of just 1.09 paths per pair.

Source-destination pairs where both are in Asia show a different behavior. For example, the proportion of AS-AS source-destination pairs within 10% of optimal jumps from 62.44% to 84.57% when going from $\kappa = 1$ to $\kappa = 2$ (for a weighted average set size of 1.31). However, achieving within 10% of optimal for close to 90% of the source-destination pairs requires $\kappa = 3$.

Note that although Table 5 shows results for $\tau = 2$ minutes for $\kappa = 2$, these values remain relatively stable for higher values of τ between 2 and 10 minutes (similar to the case of $\kappa = 1$). This implies that increasing the rate of probing does not lead to gains in latency for a significantly higher number of paths. We expand on the sensitivity of the results to τ in Section 6.3.

Interestingly, overlays designed for high performance show reduced availability as compared to the ideal situation. This is because, as illustrated in earlier examples in this chapter, better performing paths are typically constrained to share a small set of common links, leading to less path diversity and a greater vulnerability that all these shared links will simultaneously fail.

6.3 Persistence

The analysis in Section 6.2 indicates that the benefits of overlays are only mildly sensitive to the value of τ, at least in the range of 2 to 10 minutes. In this section, we explore the time sensitivity of predictive overlays by using some extreme cases. Our daily 1.5 hour samples are separated by a gap of 4 to 11 hours. We used overlays based on measurements in one 1.5 hour sample, and evaluated their performance on the next sample. While it is entirely possible that the overlay might have been suboptimal in the intervening time period, we see that around 87% of NA-NA, and 74% of AS-AS paths are within 10% of ideal even with these long term predictions. These statistics point to a high degree of consistency in the relative performance of alternative paths between a source-destination pair, for most pairs. In contrast, there is a small number of paths [20] with high short term variations, and it is difficult for a predictive overlay to optimize these paths even with κ going up to 5 or 6.

7 Future Research Directions

In this chapter, we quantified the performance and availability benefits achievable by overlay routing, and how it differs from continent to continent. The inefficiencies of the Internet have deep roots in economic considerations of the individual ISPs and are here to stay for a long time. Further, the significant geographical variations in behavior may well be artifacts of a deeper structural nature, and are not expected to even out over time as connectivity and economies improve. These facts point to a continued rapid growth in high-value traffic routed by overlay networks of CDNs. As overlay routing optimizations become more and more prevalent, the impact of these optimizations on individual ISPs operating the "underlay" and the optimizations they perform within their own networks become an interesting topic of future study [14, 19, 8].

8 Visionary Thoughts for Practitioners

After a decade of evolution, there is no doubt that CDNs now play a central role in enabling business on the Internet. Businesses in every vertical, including technology, media, entertainment, commerce, software, and government, have adopted CDN technology. The traffic hosted on CDNs continue grow by leaps and bounds, year after year. The dual challenges of enhancing the performance and availability of web sites, streaming media and applications has been a fundamental driving force of CDN evolution over the past decade. We end the chapter by refocusing our vision on those challenges and the road ahead.

- Consider that there are now retailers selling billions of dollars of goods on the Internet for whom even a 10-minute downtime of their Website during a peak period translates to millions of dollars of lost revenue and can also result in poor user perception [24]. Further, e-commerce revenue is growing at a significant rate and is expected to double every two to three years! In addition, there is growing evidence that fast downloads of Web pages are linked to larger conversion rates at e-commerce sites, leading to greater revenue. *We need to deliver content on the Internet to provide ever higher levels performance with little or no downtime.*

- Consider that there are large media and entertainment companies who rely on the Internet to disseminate content to vast numbers of end users. While they like the on-demand and ubiquitous nature of Internet streaming, they want a true television-like experience, where the video starts up immediately and never freezes! *We need to deliver content on the Internet with higher performance than traditional methods.*

- As the Internet becomes more and more entrenched as a primary source of entertainment and news, a number of content providers face the so-called flash crowd problem. *We need to deliver content on the Internet in a scalable fashion to end users even during a flash crowd, without loss of availability or performance.*

- New business trends such as outsourcing and workforce consolidation, as well as government communications necessitate exacting performance and availability standards, not just within a single country or small group of countries, but globally. It is becoming more common to have large virtual teams with individuals across the world collaborating in real-time on a single project via the Internet. Further, many novel Internet applications have more stringent performance requirements than ever. Interactive applications, such as remote shells over virtual private networks (VPNs) and multi-user games, and emerging technologies such as voice over IP (VoIP) are highly latency sensitive. *We need to meet novel and more stringent availability and performance requirements to support the next-generation of Internet applications.*

These challenges will continue to drive the field forward and shape the future CDN in the coming years.

9 Acknowledgments

The experimental results presented in this chapter appeared as a technical report in [20] and as a conference paper in [21]. Ramesh Sitaraman was supported in part by NSF Award CNS-0519894.

References

[1] Akamai Technologies, Inc. http://www.akamai.com.

[2] Akella, A., Pang, J., Maggs, B., Seshan, S., and Shaikh, A. A comparison of overlay routing and multihoming route control. In *Proc. ACM SIGCOMM*, pages 93–106, Portland, OR, Aug. 2004.

[3] Akella, A., Maggs, B., Seshan, S., Shaikh, A., and Sitaraman, R. A Measurement-Based Analysis of Multihoming. *Proceedings of the 2003 ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM)*, August 2003.

[4] Andersen, D. G. *Improving End-to-End Availability Using Overlay Networks*. PhD thesis, MIT, 2005.

[5] Andersen, D. G., Balakrishnan, H., Kaashoek, F., and Morris, R. Resilient Overlay Networks. In *18th ACM SOSP*, Banff, Canada, October 2001.

[6] Andreev, K., Maggs, B., Meyerson, A., and Sitaraman, R. Designing Overlay Multicast Networks for Streaming. *Proceedings of the Fifteenth Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA),* June 2003.

[7] CAIDA. http://www.caida.org.

[8] Clark, D., Lehr, B., Bauer, S., Faratin, P., Sami, R., and Wroclawski, J. The Growth of Internet Overlay Networks: Implications for Architecture, Industry Structure and Policy. In *33rd Research Conference on Communication, Information and Internet Policy*, Arlington, Virginia, September 2005.

[9] Cormen, T.H., Leiserson, C.E., Rivest, R.L., and Stein, C. Introduction to Algorithms. MIT Press and McGraw-Hill, 2001.

[10] Dilley, J., Maggs, B., Parikh, J., Prokop, H., Sitaraman, R., and Weihl, B. Globally distributed content delivery. *IEEE Internet Computing,* September 2002, pp. 50–58.

[11] Detour. http://www.cs.washington.edu/research/networking/detour/.

[12] Kontothanassis, L., Sitaraman, R., Wein, J., Hong, D., Kleinberg, R., Mancuso, B., Shaw, D., and, Stodolsky, D. "A Transport Layer for Live Streaming in a Content Delivery Network". *Proceedings of the IEEE, Special issue on evolution of internet technologies,* pages 1408- 1419, Volume 92, Issue 9, August 2004.

[13] Gummadi, K. P., Madhyastha, H., Gribble, S., Levy, H., and Wetherall, D. Improving the Reliability of Internet Paths with One-hop Source Routing. In *Symposium on Operating System Design and Implementation (OSDI)*, San Diego, CA, 2003.

[14] Keralapura, R., Taft, N., Chuah, C., and Iannaccone, G. Can ISPs take the heat from overlay networks? In *ACM SIGCOMM Workshop on Hot Topics in Networks (HotNets)*, 2004.

[15] Markopoulou, A., Iannaccone, G., Bhattacharyya, S., Chuah, C.-N., and Diot, C. Characterization of failures in an ip backbone. In *IEEE Infocom*, Hong Kong, 2004.

[16] NIMI. http://ncne.nlanr.net/nimi/.

[17] Paxson, V., Mahdavi, J., Adams, A., and Mathis, M. An Architecture for Large-Scale Internet Measurement. *IEEE Communications*, August 1998.

[18] PlanetLab. http://www.planet-lab.org/.

[19] Qiu, L., Yang, Y. R., Zhang, Y., and Shenker, S. On Selfish Routing in Internet-Like Environments. In *ACM SIGCOMM*, 2003.

[20] Rahul, H., Kasbekar, M., Sitaraman, R., and Berger, A. Towards Realizing the Performance and Availability Benefits of a Global Overlay Network. *MIT CSAIL TR 2005-070*, December 2005.

[21] Rahul, H., Kasbekar, M., Sitaraman, R., and Berger, A. Towards Realizing the Performance and Availability Benefits of a Global Overlay Network. *Passive and Active Measurement Conference*, Adelaide, Australia, March, 2006.

[22] RON. http://nms.csail.mit.edu/ron/.

[23] Savage, S., Collins, A., Hoffman, E., Snell, J., and Anderson, T. The End-to-End Effects of Internet Path Selection. In *Proc. ACM SIGCOMM*, pages 289–299, Cambridge, MA, 1999.

[24] Zona Research. The need for speed II. Zona Market Bulletin 5 (Apr. 2001).