Two Days in the Life of the DNS Anycast Root Servers

Ziqian Liu[2], Bradley Huffaker[1], Marina Fomenkov[1], Nevil Brownlee[3], and kc claffy[1]

[1] CAIDA, University of California at San Diego
[2] CAIDA and Beijing Jiaotong University
[3] CAIDA and The University of Auckland
{ziqian,bhuffake,marina,nevil,kc}@caida.org

Abstract. The DNS root nameservers routinely use anycast in order to improve their service to clients and increase their resilience against various types of failures. We study DNS traffic collected over a two-day period in January 2006 at anycast instances for the C, F and K root nameservers. We analyze how anycast DNS service affects the worldwide population of Internet users. To determine whether clients actually use the instance closest to them, we examine client locations for each root instance, and the geographic distances between a server and its clients. We find that frequently the choice, which is entirely determined by BGP routing, is not the geographically closest one. We also consider specific AS paths and investigate some cases where local instances have a higher than usual proportion of non-local clients. We conclude that overall, anycast roots significantly localize DNS traffic, thereby improving DNS service to clients worldwide.

Key words: DNS, anycast, Root Servers, BGP

1 Background

The Domain Name System (DNS) [1] is a fundamental component of today's Internet: it provides mappings between domain names used by people and the corresponding IP addresses required by network software. The data for this mapping is stored in a tree-structured distributed database where each nameserver is authoritative for a part of the naming tree. The DNS *root nameservers* play a vital role in the DNS as they provide authoritative referrals to nameservers for generic top-level domains (gTLD, e.g. .com, .org) and country-code top-level domains (ccTLD, e.g. .us, .cn).

When the DNS was originally designed, its global scope was not foreseen, and as a consequence of design choices had only 13 root nameservers ("roots") that would provide the bootstrap foundation for the entire DNS system. As the Internet grew beyond its birthplace in the US academic community to span the world it increasingly put pressure on this limitation, at the same time also increasing the deployment cost of any transition to a new system. Thus, anycast [2]

was presented as a solution since it would allow the system to grow beyond the static 13 instances, while avoiding a change to the existing protocol. For a DNS root nameserver, anycast provides a service whereby clients send requests to a single address and the network delivers that request to at least one, preferably the closest, server in the root nameserver's anycast group [3].

We define an *anycast group* as a set of instances that are run by the same organisation and use the same IP address, namely the *service address*, but are physically different nodes. Each instance announces (via the routing system) reachability for the same prefix/length – the so-called *service supernet* – that covers the service address and has the same origin Autonomous System (AS). The service supernet is announced from different instances by Border Gateway Protocol (BGP) such that there may be multiple competing AS paths. Instances may employ either *global* or *local* routing policy. Local instances attempt to limit their *catchment area* to their immediate peers only by announcing the service supernet with `no-export` attribute. Global instances make no such restriction, allowing BGP alone to determine their global scope, but use prepending in their AS path to decrease the likelihood of their selection over a local instance [4].

As of today, anycasting has been deployed for 6 of the 13 DNS root nameservers, namely, for the C, F, I, J, K and M roots [5]. The primary goal of using anycast was to increase the geographic diversity of the roots and isolate each region from failures in other regions; as a beneficial side effect, local populations often experience lower latency after an anycast instance is installed. As well, anycast makes it easier to increase DNS system capacity, helping protect namservers against simple DOS attacks. The expected performance gains depend on BGP making the best tradeoff between latency, path length and stability, and Internet Service Provider (ISP) cost models. BGP optimizes first ISP costs and then Autonomous System (AS) path length, attaining any gains in latency and stability as secondary effects from this optimization.

In this study we examine traffic at the anycast instances of the C, F, and K root nameservers and their client population. We substitute the geographic proximity as a proxy for latency, since latency between metropolitan areas is dominated by propagation delay [6].

2 Data

Measurements at the DNS root nameservers were conducted by the Internet Systems Consortium (ISC) and the DNS Operations and Analysis Research Center (OARC) [7] in the course of their collaboration with CAIDA. DNS-OARC provides a platform for network operators and researchers to share information and cooperate, with focus on the global DNS.

The full OARC DNS anycast dataset contains full-record `tcpdump` traces collected at the C, E, F, and K-root instances in September 2005 and January 2006. The traces mostly captured inbound traffic to each root instance, while a few instances also collected outbound traffic. For this study we selected the most complete dataset available, the "OARC Root DNS Trace Collection January 2006" [8]. It includes traces collected concurrently at all 4 C-root instances, 33

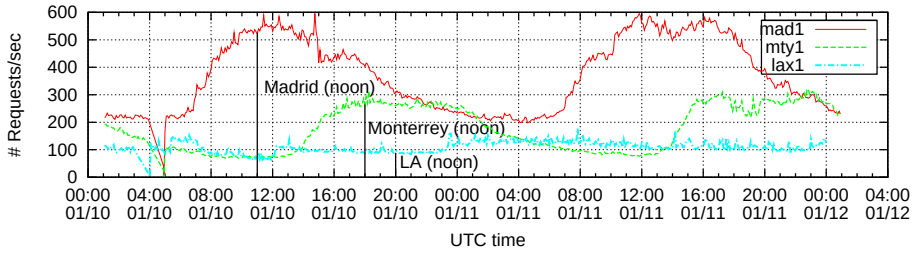


Fig. 1. Diurnal patterns of the DNS traffic to the F-root local instances `mad1` (Madrid, Spain), `mty1` (Monterrey, Mexico), and `lax1` (Los Angeles, US). For each instance, the local time noon is explicitly specified with a solid vertical line. The artifact on Jan. 10th between 4:00 and 5:00 appears because no data available for this period.

of the 37 F-root instances and 16 of the 17 K-root instances during the period from Tue Jan 10 to Wed Jan 11 2006, UTC. A common maximum interval for all measured instances is 47.2 hours or nearly two whole days.

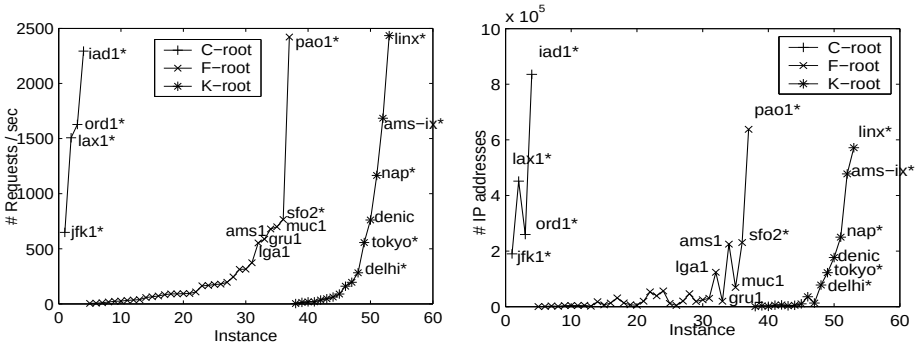
Each of the three root nameservers we measured implements a different deployment strategy [9]. All nodes of C-root are routed globally, making its topology flat. The F-root topology is hierarchical: two global nodes are geographically close, with many more widely distributed local nodes. Finally, K-root represents a case of hybrid topology with five global and 12 local nodes, all geographically distributed. The instance locations for all roots are listed in [5].

Our target data are IPv4 UDP DNS requests to each root server's anycast IP address. Some of the F and K-root instances have applicable IPv6 service addresses, and we observed a few requests destined to these addresses. Further analysis of the IPv6 DNS traffic is needed, but in this paper we focus on IPv4 traffic. We also note that for the F and K-root instances that collected TCP traffic associated with port 53, its volume was negligible, namely, ∼1.3% of total bytes and ∼3.2% of total packets.

3 Traffic Differences between Root Server Instances

3.1 Diurnal Pattern

Assuming that DNS traffic is primarily generated by humans, rather than by machines, we expect to see a clear diurnal pattern for those instances that primarily attract a client base from a small geographic area. Fig. 1 shows the time distribution of DNS requests to three F-root local instances: `mad1`, `mty1` and `lax1`. Both `mad1` and `mty1` have a clear diurnal pattern matching the local time, i.e. rising in the morning and falling towards midnight. However, `lax1` has a distinct traffic pattern, where the crest of the request curve is shifted from its local midday by ∼8 hours. This difference suggests that a large proportion of `lax1`'s requests are coming from clients who do not follow the local time of the instance, most likely, because they are located elsewhere. Indeed, as we show in



(a) Average request rate (b) Number of clients

Fig. 2. Average instance requests per second and the total number of clients. The x-axis instance order is the same in both (a) and (b). The instances are plotted in groups for C, F and K roots; within each group they are arranged in an increasing request rate order. Symbol ∗ designates global instances.

Section 4.1, although `lax1` is located in the US, ∼90% of its clients are in Asia and they generated over 70% of the total requests that this instance received.

We also studied the request time distribution of one of the global instances (not shown) and found that its curve was flatter than those of local instances. However, slight diurnal variations were still noticeable and correlated with the local time of the continent from which that global instance has the largest proportion of its clients.

3.2 Traffic Load

We characterised the traffic load of root server instances with two metrics: number of requests per second averaged over our measurement interval and total number of clients served during this interval (Fig. 2). Global instances generally have higher request rates and serve larger populations than local instances, but there is large variability in their loads. Some local instances also have fairly high traffic loads and large client populations comparable to those of the global instances. Such high loads may occur because (1) the local instance's catchment area has a high density of Internet users that generate many requests, or (2) its catchment area is topologically larger than normal. For example, the F-root local instance `ams1` is peering with AMS-IX, an Internet exchange point in Amsterdam, NL, which is one of Europe's major exchange points. Therefore, `ams1` peers with a large number of ASes via AMS-IX and attracts a higher request rate and larger number of clients than is typical for a local instance. At the same time, some local instances have extremely low load levels (less than 10 pkt/s on average over two days period), serve only a handful of clients, and are clearly underutilised.

The non-monotonically increasing curves in Fig. 2(b) indicate that the number of requests to a server can be disproportional to the number of clients it

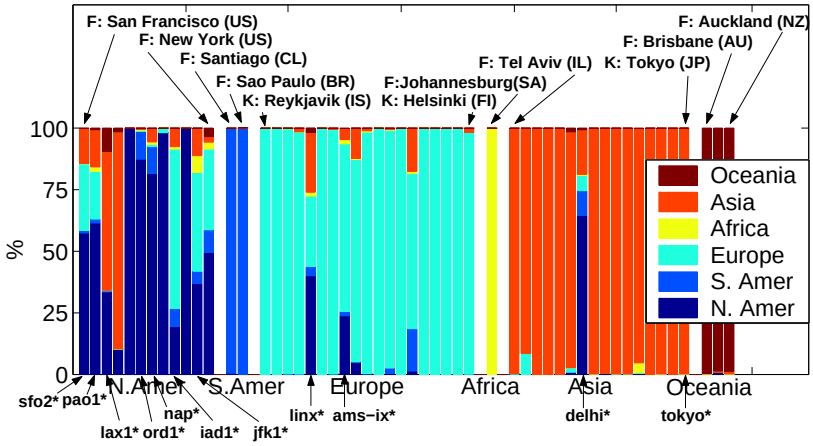


Fig. 3. Client continental distribution of instances. Each bar represents one instance, and the bars are arranged from left to right according to the instance longitude, in the west to east order. Groups delimited by white gaps represent instances located in the same continent. The anycast group (root) and the city names of the instances that are located at continent boundaries are given above the bars. Within each bar, the colored segments show the distribution of clients by continent. Global instances are marked below the bars, where the first row is for F-root, the second row is for K-root, and the third row is for C-root. To conserve space the legend overlaps some bars, but the bar color does not change within the overlapped area.

serves. Classification of users as "heavy" and "light" and a detailed analysis of their behavior patterns is a subject of future research.

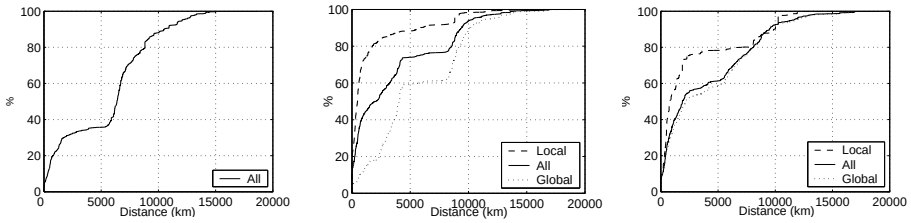
4 Anycast Coverage

4.1 Client Geographic Distribution

To discover the geographic distribution of each instance's clients, we map the client IP addresses to their geographic locations (country and continent) and coordinates (latitude and longitude) using Digital Envoy's NetAcuity database [10]. The database claims accuracy rates over 99% at the country level and 94% at the city level worldwide.

C and F-root instances are named using their corresponding airport codes, e.g. f-lax1 denotes the F-root instance at Los Angeles, while K-root instances are named either after the exchange points that support them, or their city name. Therefore, for the root server instance locations we use the coordinates of the closest airport. We then compute the geographic distance between instances and their clients as the great circle distance.

Continental Distribution. Distribution of clients by continents for each measured instance is shown in Fig. 3. Comparing clients of local and global instances we notice that clients of most global instances are indeed distributed



(a) C-root (b) F-root (c) K-root

Fig. 4. CDF of the distance from root nameservers to their clients.

worldwide. For example, the K-root global instance `linx` located at London had only 28.6% of its total clients from Europe. Others were from: North America 40%, South America 3.7%, Africa 1.6%, Asia 24.3%, and Oceania 1.8%. Most of the local instances were serving clients from the continent they are located in. For example, nearly all the clients of the F-root local instance at Santiago, Chile are from South America. Such a constrained geographical distribution is consistent with the goal of DNS anycast deployment: to provide DNS root service closer to clients.

There are exceptions among both global and local instances. Over 99.7% of K-root's `tokyo` instance were from Asia. Furthermore, 75% of its clients were less than 1000 km away, i.e. mostly in Japan. Hence, this instance behaves more like a local instance rather than a global one. The previously mentioned F-root local instance `lax1` at Los Angeles, US (the 4th bar from the left, not to be confused with the C-root global instance `lax1*` which has the same code name) has 88% of its clients from Asia, and only 10% from North America, which explains its irregular diurnal pattern in Fig. 1. Such abnormal client distributions result from the instances' BGP routing configurations, which we discuss in Section 4.2.

Distance Distribution. We also study the distance from root server instances to their clients. Fig. 4 plots, for each root server, a CDF for its local instances, its global instances, and all its instances combined Only one curve is given for the C-root instances since they are all global.

Fig. 4 shows that the majority of the local instances were serving clients who are geographically close to them – 80% of the F-root local instances' clients and 70% of the K-root local instances' clients were within 1800 km. Distances between the global instances and their clients are generally longer, e.g. for C-root, over 60% of the clients were beyond 5000 km, and F- and K-root both had 40% of their clients beyond 5000 km. The F and K roots had lower proportions of clients who were far away from their servers because these anycast groups include multiple local instances all over the world while the C-root group currently has only 4 instances and they are all global.

Flat segments in the CDF curves (around 5000 km for the C and K roots, and the especially prominent one from 5000 to 8000 km for the F-root) approximately correspond to the distances across the Atlantic Ocean (from North America

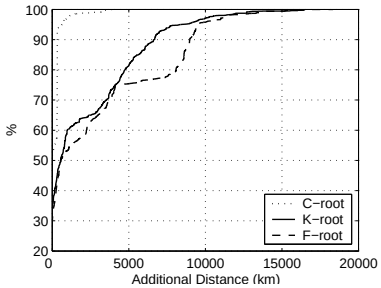


Fig. 5. CDF of additional distance travelled by requests to instances.

to Europe) and the Pacific Ocean (from North America to Asia), respectively. Obviously, fewer clients are found in the ocean areas.

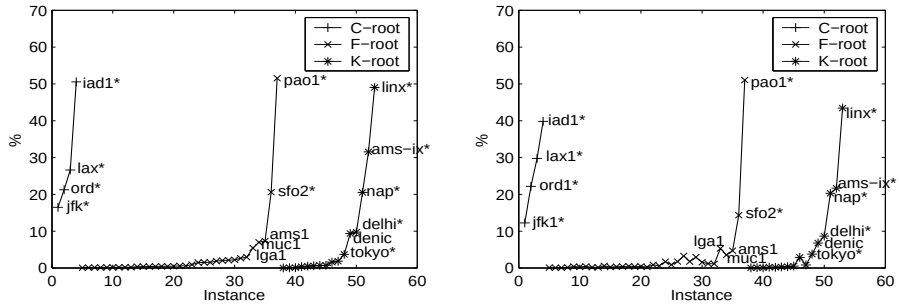
Additional Distance. We wanted to investigate whether the BGP always chooses the instance with the lowest delay. For this analysis, we use the geographic proximity as a proxy for latency. A comprehensive study [6] shows that, geographic distance usually correlates well with minimum network delay. Later studies [11, 12] also used geographic distance to compute network delay.

For a given client, we define the *serving instance* as the instance the client actually uses, and the *optimal instance* as the geographically closest instance from the same anycast group. We ignore the tiny number of clients that sent requests to more than one instance. (see Section 4.3 below). We then define the client's *additional distance* as the distance to its serving instance minus the distance to its optimal instance. An additional distance of zero indicates that the client queried an optimal instance while a positive value suggests a possible improvement.

Analyzing the CDF of the additional distance (Fig. 5), we saw that 52% of C-root's clients were served by their optimal C-root instance, and another 40% had short additional distances. This optimised selection is due to the flat topology of the C-root anycast group, i.e., all instances are global. In contrast, only 35% of F-root's clients and only 29% of the K-root's clients were served by their optimal instances. Given that the speed of light in fiber is about 2×10^8m/s, an additional 5000 km of geographical distance adds a 25 ms delay. Our results imply that a significant number of clients would benefit if routing configurations of their local DNS root instances were optimized to route these clients to their optimal instance, thereby reducing their DNS service delay.

4.2 Topological Coverage

We studied the topological coverage of the Internet by anycast clouds of the C, F, and K root nameservers. Using the RouteViews BGP tables [13] collected on 10 Jan 2006, we mapped each client IP address in our data to its corresponding prefix by longest matching, and so determined its origin AS. Out of 21883 ASes seen in RouteViews tables on that day, we observed IP addresses belonging to 19237 ASes (∼88%) among our clients.



(a) AS level (b) Prefix level

Fig. 6. Topological scope of the instances. The x-axis instance order is the same in (a) and (b). The instances are plotted in groups for C, F and K roots; within each group they are arranged in an increasing AS coverage percentage order. The percentage shown for of each instance is the number of ASes/prefixes seen by the given instance divided by the total number of ASes/prefixes seen by all three roots combined.

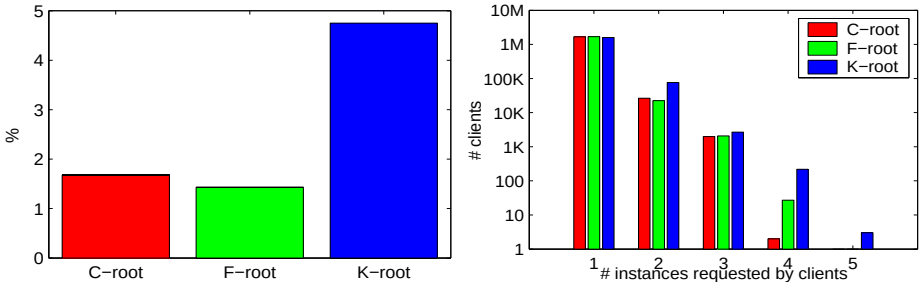
Fig. 6 shows both the AS-level and prefix-level coverage for each instance relative to the total number of ASes (prefixes) seen by all instances of the three root nameservers. As expected, most of the global instances have much higher topology coverage than the local instances.

Two exceptions are: (1) the K-root local instance `denic` in Frankfurt, Germany had a wider topological scope than any other local instances; (2) the K-root global instance `tokyo` saw a rather small fraction of ASes and prefixes. Such exceptions can be explained by RouteViews BGP data.

Knowing the IP address of the K-root anycast service supernet and using the AS peering information published at the root server website [14], we extracted the AS paths to each of its instances. One of the three observed AS paths to `denic` is `12956 8763 25152`. According to the RIPE-NCC *whois* database, AS12956, belongs to Telefonica, which has a global network infrastructure. The presence of this path explains why `denic` has a high topological coverage, and, correspondingly, a high traffic load and a large number of clients (cf. Fig. 2).

Considering the K-root instance `tokyo`, we note that the global instances used AS-path prepending to intentionally lengthen their paths. This instance announced a triple AS-prepended path, i.e. `4713 25152 25152 25152 25152` which was the longest among all of the five K-root global instances. Such a long AS path caused `tokyo` to be seldom chosen by BGP for global clients who sent queries to the K-root server. Therefore, the clients of `tokyo` were mostly local (cf. Fig. 3).

Finally, we saw in Section 4.1 that the F-root local instance `lax1` had most of its clients coming from Asia. One of the three AS paths we observe to this instance was: `7660 2516 27318 3557`, where AS7660 and AS2516 are in Japan thus explaining the source of the Asian clients.



(a) Percentage of the clients switching instances

(b) # clients (log scale) vs # instances queried within same anycast group

Fig. 7. Instance switching within the same anycast group.

4.3 Instance Affinity

Anycast improves stability by shortening AS paths, thus decreasing the number of possible failure points. However, this enhancement comes at the cost of increased chance of inconsistency among instances [2] and of clients' transparent shifting to different instances. As long as DNS traffic is dominated by UDP packets, this route flapping is unimportant, but it may pose a serious problem if stateful transactions such as TCP or multiple fragmented UDP packets become more prominent [15]. Fortunately, recent studies [9, 16, 17] suggest that the impact of routing switches on the query performance is rather minimal.

We observed that a small fraction of clients did switch instances during the two days: 1.7% of the C-root clients, 1.4% of the F-root clients, and 4.7% of the K-root clients (Fig. 7(a)). These percentages correlate with the number of global instances each root server has (4 for C-root, 2 for F-root, and 5 for K-root), since the clients of a global instance are more easily affected by routing fluctuations. Actually, the two F-root global instances together saw approximately 99.8% of the total clients who switched F-root instances, and the five K-root global instances together saw 86% of the total clients who switched K-root instances.

Fig. 7(b) shows how many clients queried how many instances. Focusing on the clients who used the most instances, we found that the two C-root clients who requested four instances were from Brazil and Bolivia and the three K-root clients who requested five instances were all from Uruguay. Note that neither the C-root nor the K-root had an instance in South America. For F-root, the 27 clients who requested four instances were all from the UK where the F-root has a local instance `lcy1`, but the catchment area of this instance was limited. Actually, those 27 clients never requested from `lcy1`, but switched between `ams1`, `lga1`, `pao1`, and `sfo2`. A detailed analysis of unstable clients could help network designers decide where to place new instances.

5 Conclusion

From the diurnal patterns of request rates and from the observed geographic clustering of clients around instances we conclude that the current method for

limiting the catchment areas of local instances appears to be generally successful. A few exceptions, such as the F-root local instance `lax1` or the K-root local instance `denic`, drew their clients from further away regions due to peculiar routing configurations.

Instance selection by BGP is highly stable. Over a two-day period less than 2% of both C-root and F-root clients and <5% of K-root clients experienced an instance change. Since UDP connections are stateless, the vast majority of clients would not be harmed by such changes, apart from the unavoidable delay created by BGP convergence. Although the instance flapping could be problematic to TCP's stateful connections [15], in our data sample TCP packets constituted only 3.2% of all DNS root packets.

Overall, the transition to anycasting by the DNS root nameservers not only extended the origial design limit of 13 DNS roots, but it also provides increased capacity and resilience, thereby improving DNS service worldwide.

Acknowledgements. We thank ISC, RIPE, and Cogent for collecting the datasets used in this study. P. Vixie, K. Mitch, and B. Watson of ISC helped with data storage and answered questions on F-root's anycast deployment. A. Robachevsky and C. Coltekin from RIPE provided feedback on K-root's anycast deployment. This work was supported by NSF Grant OCI-0427144.

References

1. Mockapetris, P.: Domain Names - Concepts and Facilities, Internet Standard 0013 (RFC 1034, 1035), Nov. 1987.
2. Hardie, T.: Distributing Authoritative Nameservers via Shared Unicast Addresses. RFC 3258, Apr. 2002.
3. Partridge, C., Mendez, T., Milliken, W.: Host Anycasting Service. RFC 1546, 1993.
4. Abley, J.:Hierachical Anycast for Global Service Distribution. `http://www.isc.org/pubs/tn/isc-tn-2003-1.html`
5. DNS root nameservers web sites. `http://www.root-servers.org/`
6. Padmanabhan, V.N., Subramanian, L.: An Investigation of Geographic mapping techniques for Internet Hosts. ACM SIGCOMM, Aug. 2001.
7. OARC. `https://oarc.isc.org/docs/dns-oarc-overview.html`
8. OARC Root DNS Trace Collection January 2006 `http://imdc.datcat.org/collection/1-00BC-Z=OARC+Root+DNS+January+2006`
9. Colitti, L., Romijn, E., Uijterwaal, H., Robachevsky, A.: Evaluating The Effect of Anycast on DNS root nameservers. Unpublished paper, Jul. 2006.
10. NetAcuity. `http://www.digital-element.net`
11. Spring, N., Mahajan, R., Anderson, T.: Quantifying the Causes of Path Inflation. ACM SIGCOMM, Aug. 2003.
12. Sarat, S., Pappas, V., Terzis, A.: On the Use of Anycast in DNS. ACM SIGMETRICS, Jun. 2005
13. Route Views Project. `http://www.routeviews.org`
14. K-root Homepage. `http://k.root-servers.org/`
15. Barber, P., Larson, M., Kosters, M., Toscano, P.: Life and Times of J-root. NANOG 32, Oct. 2004
16. Boothe, P., Bush, R.: DNS Anycast Stability.19th APNIC, Feb. 2005.
17. Karrenberg, D.:Anycast and BGP Stability.34th NANOG, May 2005.