Understanding IP Anycast

Hitesh Ballani, Paul Francis
Cornell University, Ithaca, NY

I. Introduction

Anycast is a paradigm for communicating with one member of a group. An anycast service, when implemented at the network layer, is called *Network-layer* or simply *IP Anycast*. IP anycast delivers packets destined to an anycast address to a member of the anycast group, typically the one which is closest to the sender in terms of the metrics used by the routing protocol.

Since Partridge et. al. [1] proposed IP anycast, there has been a lot of work regarding the use of anycast to provide robust and efficient service discovery [2] [3] and in other applications such as DDoS sinkholes [4]. Inter-domain Anycast has been used for the anycasting of the DNS root-servers [5] [6] and AS-112 servers [7]. A number of proposals have also sought to address the problems afflicting IP anycast[1] [8] [9] [10].

In spite of the use of anycast in critical infrastructural services, IP Anycast and its interaction with existing elements of the Internet (such as inter-domain routing) is not well understood. So, a study of anycast has use in this context as the root-server deployment is important and growing. More generally, determining the performance of IP anycast is necessary before we can comment on the feasibility of the IP anycast (and other aforementioned proposals) and its applications.

In this paper we present the first detailed analysis of IP anycast as used in the anycasting of the root-servers. The main results of our study are:

- The anycasting of an IP prefix does not have any unfavorable interactions with the routing system. Hence, IP anycast offers very good affinity[2] - this alleviates concerns regarding running connection oriented services on top of anycast.
- IP Anycast, by itself, does not offer proximity in terms of metrics such as latency. IP Anycast's backwards compatibility derives from the fact that it is transparent to existing routing protocols, but this transparency also implies that in many cases inter-domain routing, which was designed with unicast path-selection in mind, chooses anycast locations which are not close to the source. We also present deployment schemes that might allow anycast to achieve good latency based proximity.

II. Teminology

In this section we define some terms which are used frequently used in the rest of the paper:

- Anycast Address : the anycast group name; given that we are dealing with IP Anycast, it is the IP address by which the group members can be reached.
- Client : the host which is trying to access an anycast group by sending packets to the anycast address.
- Anycast Location : when a client sends a packet to the anycast address, the packet reaches one of the locations where members of the anycast group are located; such a location is referred to as an anycast location.
- Degree of anycasting : the number of locations for a given anycast group.

III. Affinity

IP Anycast, being an IP layer service, provides best effort delivery of anycast packets to a group member. However, the fact that it is a network layer service implies that two consecutive packets from a single client need not be delivered to the same member. Such an occurrence, hereon referred to as a flap, casts doubts on efforts to run connection oriented services on top of IP Anycast. Hence, determining the affinity offered by IP anycast is important for evaluating the quality of IP anycast as a substrate for connection oriented services.

The affinity observed by a client, when accessing an anycast group, can be poor due to the following reasons:

- Client is multi-homed : in such a scenario a switch across the immediately upstream provider being used might cause a flap. While the traditional use of multihoming involves load-balancing at coarse granularity or in a stateful fashion, modern techniques for performance improvement might involve fast switches.
- One (or more) anycast locations are unstable : such a scenario would lead to frequent BGP events for the anycast prefix. This might appear to have the same impact as a unicast prefix whose origin is unstable and hence, is frequently unreachable. But the fact that the anycast prefix is advertised from a number of places makes the situation worse. And route flap-dampening by routers implies that extreme instability would have a severe impact on the reachability for the anycast prefix.

[1] the problems include scalability by the number of anycast groups, difficulty of deployment etc.; these have restricted the use of IP anycast to critical infrastructure services

[2] tendency of subsequent packets of a ”connection” to be delivered to the same target

Anycast-Server	Degree
c-root	4
f-root	28
i-root	19
j-root	15
k-root	11
m-root	3
as-112	20

TABLE I

ANYCAST SERVERS AND THEIR DEGREES

Continent/Country	PL Nodes	Traceroute-servers
Africa	0	3
Asia	22	26
Australia	3	12
S.America	1	8
Canada	12	1
Europe	31	152
US	94	42
Total	163	244

TABLE II

GEOGRAPHIC SPREAD FOR PL NODES AND TRACEROUTE SERVERS

- The interaction of anycasting a prefix with inter-domain routing mechanisms : the fact that a prefix is anycasted means the routers will have more routes to the prefix available to them. This might have unfavorable interaction with the existing routing system - the routers might switch between different paths to the same prefix resulting in frequent flaps. Note, the fact that an anycasted AS with x locations is the same as a multihomed AS with x upstream providers at the inter-domain level does not answer the question raised here because we have not had the experience with AS'es with ∼30 upstream providers as is the case with some of the anycast deployments.

In the following sections we present experiments and analysis done to determine the affinity offered by some of the existing anycast deployments and in cases where the affinity is poor, we try to nail it down to one of the aforementioned causes.

A. Data Collection

In this section, we describe the traces collected by us in order to answer the questions raised above:

- **PL-3-mon** : The experiment involved probing 7 anycast destinations (6 anycasted DNS root-servers and the AS-112 servers) from 163 Planetlab [11] sites for 3 months (Dec'04-Mar'05). The number of locations associated with each anycast destination are detailed in table I while the geographical spread of the Planetlab sites is shown in table II. Each location of each of these anycast destinations has been configured by their operators to answer DNS TXT queries addressed to the anycast address with the location of the box answering it [12]. So a client can determine the particular location of an anycast group it is accessing by sending TXT type queries to the anycast address.

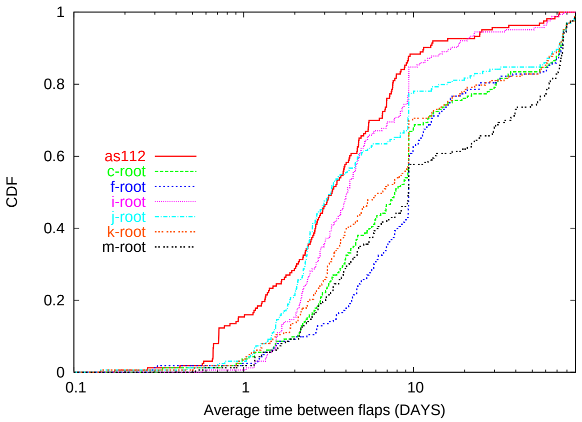

Fig. 1. CDF for inter-flap interval as seen at Planetlab nodes using the PL-3-mon trace

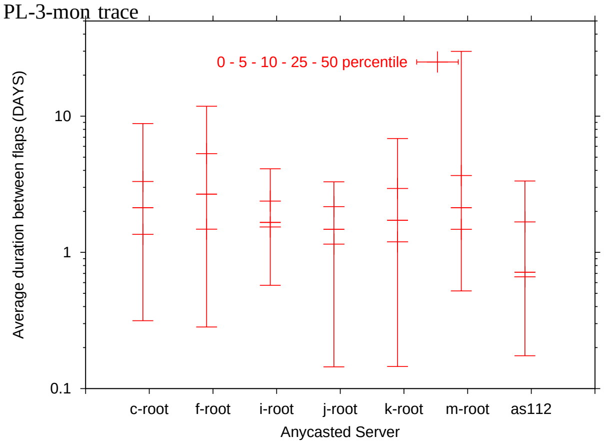

Fig. 2. Percentiles for inter-flap interval as seen at Planetlab nodes using the PL-3-mon trace

For a given anycast destination, each of our probe-points queries the destination once every 10 seconds and logs the location it reaches. The probes timeout after 5 seconds prompting another immediate probe; timeout of three consecutive probes is logged as a FAILURE event.

- **Server-trace** : In this experiment, we conducted traceroutes from 244 publicly accessible traceroute-servers to 4 anycast destinations for a week each; the geographic spread of these servers is given in table II. Given the restrictions on the load that can be acceptably placed on these, each destination was traced to in a separate week. The load restriction also led us to choose a 60 second interval between the start of two consecutive traceroutes.

B. Measurement results

This section presents analysis of the data-sets described in the previous section:

- The *PL-3-mon* trace provides the location reached by a probe point for each probe to an anycast destination and hence, allows us to observe the flaps experienced by the probe points. We used this to calculate the average inter-flap interval for every probe point and anycast destination pair. The **average inter-flap interval** is obtained by averaging the total probe duration across the number of flaps observed when probing the anycast destination from the probe point and hence, is an indicator of the stability of the anycast destination as seen from the probe point.

The CDF for the average inter-flap interval is shown in figure 1. Figure 2 discretizes the CDF and shows various percentiles for the average inter-flap interval. These figures show that majority of the probe points observed less than a flap per day for all the anycast destinations. The measurements reveal that the probability that a two minute connection breaks due to a flap is about 1 in 3500[3,4]. Also, probe points with a small inter-flap interval had their average skewed by small periods of instability when they observed lots of flaps.

- While the *server-trace* has traceroutes to the anycast destinations at an interval of 60 seconds, it is not clear if these probes are frequent enough to not miss short-duration flaps. If there are two successive flaps within a period of x seconds, then probing at an interval of x seconds might miss one (or both) of the flaps. We looked at the typical time between flaps in order to determine the rate of probing needed to capture most of the flaps. The analysis of the *PL-3-mon* trace showed that with a probing rate of once a minute, we would have missed less than 5% of the flaps in all the sets (i.e. across all the anycast destinations). The lack of short duration flaps seems to agree with the current nature of inter-domain routing[5]; a failure might lead to a number of flaps, but it is very rare to have a lot of very closely spaced flaps. Hence, probing the anycast destination at once a minute should suffice to capture almost all the flap activity.
- The lack of short duration flaps allows for the use of traceroutes collected in the *server-trace* to look for flaps in the anycast destinations as seen from the traceroute-servers. These servers buttress our argument by providing a set of geographically diverse probe points which is not as biased as the PL deployment. Also, in terms of characterization of the root-server anycast deployment, these servers do seem to be representative of the DNS servers provided by ISPs to their clients.

  Based on the periodic traceroutes we were able to determine the flaps experienced by the traceroute-servers when probing the anycast destinations and hence, calculated the average inter-flap interval for every probe point and anycast destination pair. Figure 3 shows the CDF for the average inter-flap duration[6]. Overall, the traceroute-servers too seem to have a stable view of the anycast destinations with more than 95% of the nodes experiencing flaps at a frequency of less than one per day.

  A closer look at the CDF shows that one traceroute-server did observe flaps at a high rate (∼20 minutes). We looked at the probes from this server and found that the site hosting the server was load balancing its traffic across multiple upstream ISPs at a high rate and hence the frequent flaps. This means that as expected, load balancing across multiple upstream ISPs at a high rate[7] has a deleterious impact on the anycast flaps. However, in our study we did not find many instances of sites using such load-balancing approaches. While this can be attributed to the kind of sites at our disposal, we can conclude that the feasibility of IP anycast to support connection oriented services might just boil down the popularity and necessity of such load-balancing across upstream ISPs.

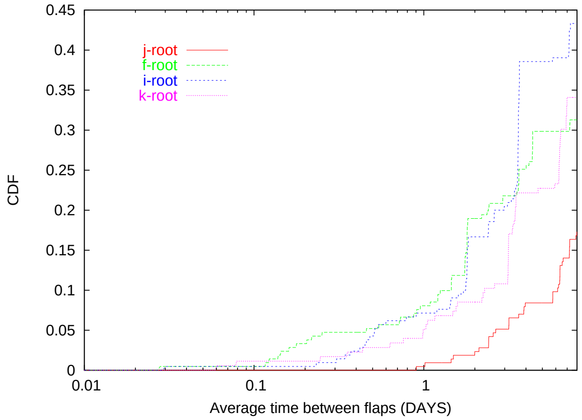

Fig. 3. CDF for flaps as seen from traceroute servers

C. BGP-level analysis

The active probing experiments described in the previous section show that the root-server anycast deployment offers very good affinity. In this section, we look at BGP level activity for the anycast prefixes as contrasted against the activity for unicast prefixes in the Internet. The comparison would reveal what impact, if any, does the anycasting of a prefix have on the prefix dynamics as seen at the inter-domain routing level.

- We used the BGP updates collected by publicly available BGP repositories (Route-Views [13] and

[3] note that this figure is extremely conservative as it assumes a uniform distribution of flaps across time and across different probe points which is certainly not the case in practice

[4] we presume that long duration connections would rely on anycast for discovery and be redirected to use unicast once the discovery is made; hence, the flaps should not be a concern even for long running connections

[5] for eg, the MinRouteAdvertisementTimer controls the rate of updates for each destination and is typically set to 30 seconds

[6] note that the *server-trace* involved probing to each destination for a week each

[7] stateless load balancing or in the worst case scenario, per packet load balancing; stateful load balancing would not lead to such problems

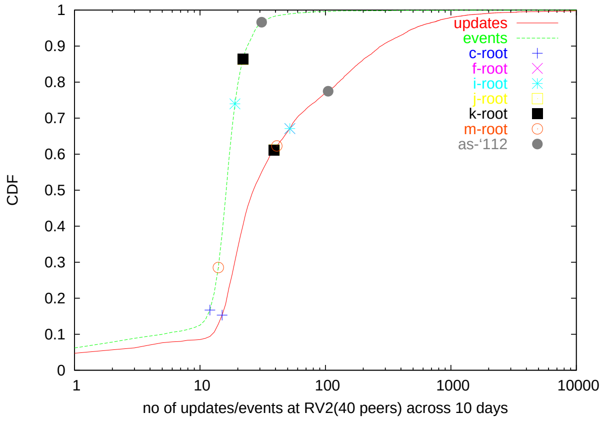

Fig. 4. CDF for BGP updates and events seen at Route-Views (rv2): the anycast prefixes fall in the highly stable range of prefixes with ∼20 events across the 10 days observed

RIPE [14]) to characterize and contrast BGP dynamics for the anycast prefix against unicast prefixes. These repositories have an EBGP peering with a large number of autonomous systems (ASs) and hence, the BGP updates for a particular prefix seen at the repositories provide information about any changes in the best route to the prefix used by the participating ASs.

- The BGP update data available at the repositories has a few anomalies that must be accounted for before it can be used for analysis. This includes session resets at the collection boxes, redundant updates sent by the routers. The data was pre-processed to account for these anomalies – a discussion of the cause for the anomalies and the pre-processing can be found in [15]. The analysis below also uses the *event* abstraction: an event is defined as a collection of BGP updates such that two consecutive updates are separated by no more than 120 seconds. The use of *events* for the analysis ensures that we avoid some of the timing effects that govern the number of routing updates generated by a particular inter-domain activity.
- Given the BGP updates for all prefixes present in the BGP table at the repositories, we plotted the CDF for the updates and events for each prefix. We then marked out the position of the anycast prefixes along the CDF curve. Figures 4 and 5 show the CDFs observed at Route-Views (route-views2) and RIPE (rrc00) respectively.
- While the anycast prefixes seem to have relatively highly activity, they are pretty stable in absolute terms. Rexford et. al. had argued that a large proportion of prefixes tend to be highly stable - and the anycast prefixes appear to fall in this range. This buttresses our argument regarding the stability of anycast destinations - the anycasting of a prefix (at least to a degree of ∼30) does not interact badly with the existing routing infrastructure.
- We repeated the above analysis for BGP updates collected at other route-collectors of Route-Views and RIPE; the results were similar.

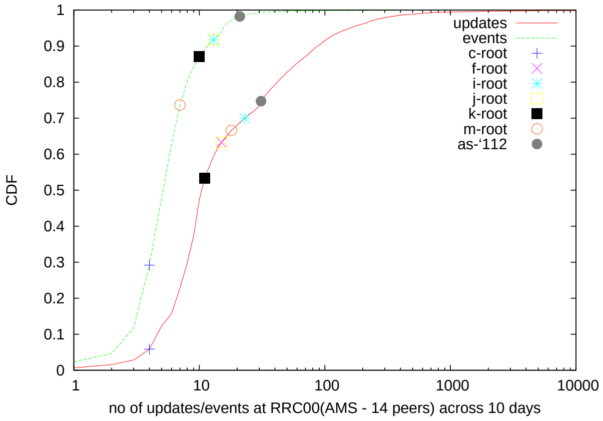

Fig. 5. CDF for BGP updates seen at RIPE RRC-00 across 10 days

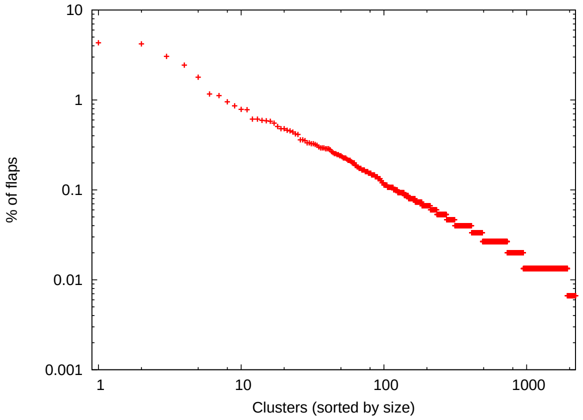

Fig. 6. Clustering of flaps

D. Flap-cause analysis

We also tried to determine the cause of the flaps observed in the *PL-3-mon* data-set - are the observed flaps a result of routers switching paths due to the greater number of options or due to events such as links/anycast-destinations failing and so on.

- We clustered the flaps for all sources and all destinations across time, i.e. a flap seen within 120 seconds of the last flap was part of the same cluster. The ∼16000 flaps were clustered into 2163 clusters. The relative contribution of these clusters towards the total number of flaps is shown in figure 6. As can be seen, a very small number of clusters contribute a large fraction of the flaps - the 20 biggest clusters (1% of the clusters) account for 27% of the flaps.
- While the clusters in general do not represent BGP events, we believe that large clusters are the result of major routing events. We looked at large clusters and found sufficient evidence of this. For example, the largest cluster comprising of 650 flaps seemed to be due to a routing event in the Stockholm area. Figure

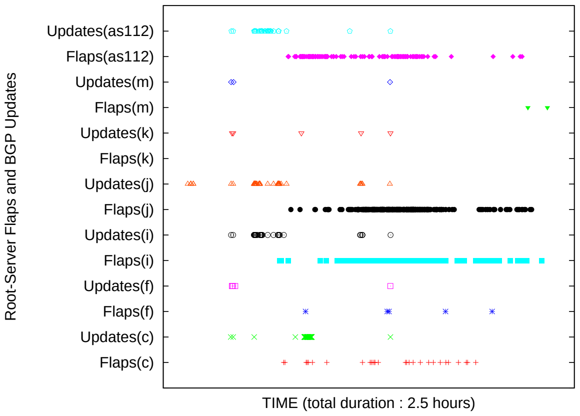

Fig. 7. Updates and flaps during the largest cluster

7 shows the BGP updates and the flaps seen during this period. As can be seen there is good correlation between the flurry of BGP updates and the subsequent flaps (note the updates and flaps for as112, j-root and i-root, all of whom have Stockholm as one of the anycast locations).

E. Summary

We now summarize the results of our experiments regarding the affinity offered current inter-domain anycast deployments:

- Our active measurements showed the measured anycast deployments offer very good affinity and this observation is corroborated by BGP-level stability analysis for the anycast prefixes. As a matter of fact, many of the observed flaps seem to be the result of routing events which shows up in the number of affected probe points.
- Overall, our observations suggest that the anycasting of a prefix, at least to a degree of 30, does not have harmful interactions with the existing routing system. Hence, inter-domain anycast provides a substrate which is stable enough to support connection oriented services on top.
- The few cases where anycast does perform badly involve multihoming at the client (with fast load-balancing across the upstream providers) – if such techniques were to become popular amongst ISPs (which did not come across in our study), native IP-Anycast may not remain suitable for offering connection oriented services.

IV. Proximity

Part of the value of an anycast service as a server selection primitive is its ability to find close-by members of the anycast group. And IP Anycast, by its very nature, delivers client packets to the anycast group member which is closest to the client in terms of metrics used by the routing protocol; for a inter-domain anycast group, this is the metric used by BGP for route selection. Hence, an important question is if this proximity in terms of routing protocol metrics leads to proximity in terms of metrics such a distance/latency. In this section we take a look at this question.

- In order to determine the quality of proximity offered by a given native IP Anycast deployment to a particular client, we need to determine the following latencies from the client:
  - The latency of each location of the anycast destination - this is the *unicast latency* from the client to the particular location and can be determined by probing the unicast address of the location from the client.
  - The latency of the anycast address associated with the destination - this is the *anycast latency* from the client to the destination and can be determined by probing the anycast address from the client.

  For anycast to offer good latency-based proximity, the anycast latency should be close to the minimum unicast latency, i.e. anycast packets should go the location which is closest.
- While determining the *anycast* and *unicast latencies* by active probing would restrict us to measuring latencies from clients under our control, we chose to use King [16] for the same. The King approach allows measurement of the latency between arbitrary Internet end hosts. It does so by using recursive DNS queries to determine the latency between the authoritative name servers for the end-hosts; this gives an estimate of the latency between the actual end hosts.
- While the basic King approach suffices for determining the *unicast latencies* from the client to all locations of an anycast destination, determining the *anycast latency* for the client requires extra effort (see Section II.C [16] for more details). For example, using King to determine the latency between a client C and *f.root-servers.net* would not yield the anycast latency from C to f-root, rather such a query would give the latency between C and any one of the nameservers which are authoritative for *root-servers.net* (i.e. any one of the 13 root nameservers). For measuring the anycast latency from the client C to f-root, we need to ”trick” C (or C’s authoritative name-server) into querying f-root directly and this can be achieved by using a lame-delegation to point to f-root. We used a domain name under our control (*xyz.cc*) to create a lame delegation from *f-root.xyz.cc* to *f.root-servers.net*. Hence, a recursive query for *random-number.f-root.xyz.cc* to C’s nameserver tricked it into contacting f.root-servers.net, thus allowing us to measure the anycast latency from C to the f-root server.
- A pre-requisite for performing the proximity experiment for a given anycast destination is knowledge of the unicast IP addresses for all the locations of the particular destination. Hence, we performed the

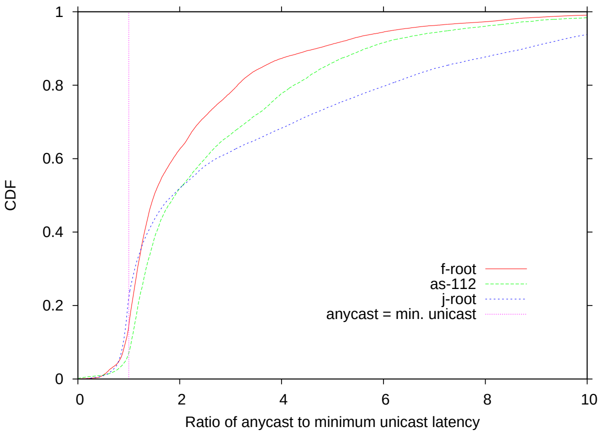

Fig. 8. CDF for ”anycast latency” to ”minimum unicast latency” ratio for j-root, f-root and the as-112 servers

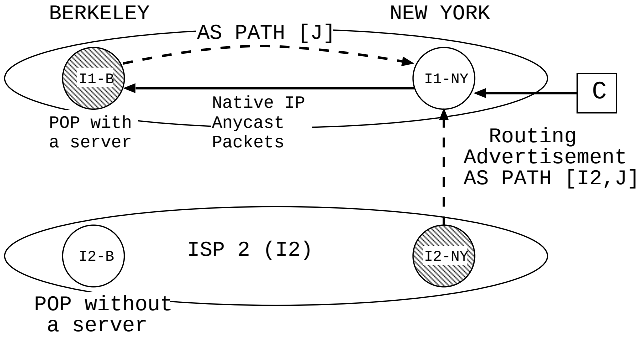

Fig. 9. Native IP anycast inefficiency - packets from client C in New York destined to the native IP anycast address are routed to the anycast server in Berkeley, even though there is a server in New York

experiment for three anycast destinations[8]. We were able to determine the *anycast* and *unicast latencies* to these destinations from ∼30000 clients. The CDF for the *anycast latency* to the *minimum unicast latency* ratio is plotted in figure 8. As can be seen from the figure, the three anycast deployments measured in this experiment do not provide good latency-based selection[9].

- We believe the inefficacy of anycast when selecting close-by root-servers might be due to the way the root-servers have been deployed; for example, in case of the j-root servers, all 15 anycasted servers are placed in POPs of different ISPs. A possible problem with this approach is illustrated in figure 9. The figure shows 2 ISP networks- I1 and I2, each having a POP in New York and in Berkeley. It also shows a native IP anycast deployment (AS number J) with two servers - one hosted at the New York POP of I2 (I2-NY) and the other at the Berkeley POP of I1 (I1-B). The figure has these POPs highlighted. The anycast servers have an EBGP relation with the routers of the hosting POP; hence, the anycast prefix is advertised with J as the origin AS. Now, if a client (C) in the New York area sends packets to the anycast address and these reach POP I1-NY, they will be routed to the server hosted at I1-B. This is because the routers in I1-NY would prefer the 1 AS-hop path ([J]) through I1-B to the anycasted server over the 2 AS-hop path ([I2,J]) through I2-NY. Note that the anycasted server hosted at I1-B represents a customer of I1 and so, it would be very uncommon for I1 to steer these packets towards I2-NY due to local policies (local preference values); rather the AS path length would dictate the path.

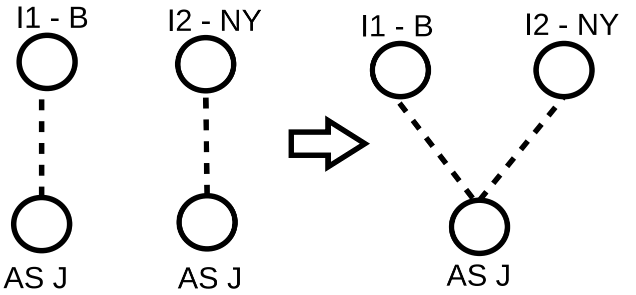

Fig. 10. Equivalence between an anycasted autonomous system and a multi-homed autonomous system from the perspective of inter-domain routing; however, AS J seems to be multi-homed to two far apart locations

- On a more general note, from the point of view of inter-domain routing, an anycasted autonomous system is equivalent to a multi-homed autonomous system (see figure 10). However, anycasting introduces multi-homing scenarios which differ significantly from normal multi-homing scenarios. For example, figure 11 shows two common multihoming scenarios involving an autonomous system (AS J) multihomed to two providers - AS I1 and AS I2. However, in both the scenarios, AS J has peerings to POPs of I1 and I2 which are close by (either in New York or in Berkeley). Hence, the AS-hop based path selection used by inter-domain routing does an acceptable job of selecting paths from clients to destinations in AS J. However, an anycasted autonomous system (AS J in figure 10) will typically have peerings with POPs of upstream ISPs which are not close by (such as I1-B and I2-NY) and so, path selection to destinations in AS J based on number of AS-hops has a much higher chance of making an unsuitable choice.
- Although negative, the importance of the result cannot be overemphasized. It brings to light the fact that while the routing protocols used to choose paths to unicast destinations work naturally for anycast destinations[10] too, the metrics used for routing decisions can lead to a poor choice of the anycast location. While changing

[8] the unicast IP addresses for the rest of the root-servers were not available to us

[9] the poor selection in case of f-root server and the as112 servers can be attributed to their use of hierarchical anycast [6]; however, this is not the case for the j-root server

[10] this is why IP Anycast is backwards compatible with no changes in routers or routing protocols

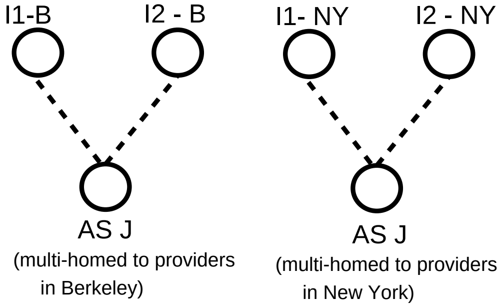

Fig. 11. Typical Multi-homing scenarios : an autonomous system multihomed to close-by POPs of upstream providers

the routing protocols to differentiate between anycast and unicast prefixes would be one possible approach to address this, a more practical approach would involve being selective about the locations where anycast servers are placed. For example, one such deployment approach would involve ensuring an ISP that hosts anycast servers is sufficiently covered, i.e., there should be anycast servers at many POPs of the ISP. For example, deployment of the two servers in figure 9 at both of the POPs of I1 (I1-NY and I1-B) or I2 (I2-NY and I2-B) would avoid the problem of long paths. We believe that such an approach would ensure that current inter-domain routing will be able to find paths to close-by anycast locations while maintaining the backwards compatibility of IP Anycast.

REFERENCES

[1] C. Partridge, T. Mendez, and W. Milliken, “RFC 1546 - Host Anycasting Service,” November 1993.

[2] E. Basturk, R. Haas, R. Engel, D. Kandlur, V. Peris, and D. Saha, “Using IP Anycast For Load Distribution And Server Location,” in *Proc. of IEEE Globecom Global Internet Mini Conference*, November 1998.

[3] S. Matsunaga, S. Ata, H. Kitamura, and M. Murata, “Applications of IPv6 Anycasting,” draft-ata-ipv6-anycast-app-00, February 2005.

[4] B. Greene and D. McPherson, “ISP Security: Deploying and Using Sinkholes,” www.nanog.org/mtg-0306/sink.html, June 2003, NANOG TALK.

[5] T. Hardy, “RFC 3258 - Distributing Authoritative Name Servers via Shared Unicast Addresses,” April 2002.

[6] J. Abley, “Hierarchical Anycast for Global Service Distribution,” ISC Technical Note ISC-TN-2003-1 www.isc.org/tn/isc-tn-2003-1.html.

[7] “AS112 Project Home Page,” May 2006, www.as112.net.

[8] D. Katabi and J. Wroclawski, “A framework for scalable global IP-anycast (GIA),” in *Proc. of ACM SIGCOMM*, 2000.

[9] H. Ballani and P. Francis, “Towards a global IP Anycast service,” in *Proc. of ACM SIGCOMM*, August 2005.

[10] I. Stoica, D. Adkins, S. Zhuang, S. Shenker, and S. Surana, “Internet Indirection Infrastructure,” in *Proc. of ACM SIGCOMM*, 2002.

[11] B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman, “PlanetLab: An Overlay Testbed for Broad-Coverage Services,” *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 3, pp. 3–12, July 2003.

[12] “ISC F-Root Sites,” www.isc.org/index.pl?/ops/f-root/.

[13] “Route-Views,” www.route-views.org.

[14] “RIPE RIS,” http://www.ripe.net/projects/ris/rawdata.html.

[15] J. Rexford, J. Wang, Z. Xiao, and Y. Zhang, “BGP routing stability of popular destinations,” in *Proc. of the 2nd ACM SIGCOMM Workshop on Internet measurment (IMW’02)*, 2002.

[16] K. P. Gummadi, S. Saroiu, and S. D. Gribble, “King: Estimating Latency between Arbitrary Internet End Hosts,” in *Proc. of the SIGCOMM Internet Measurement Workshop*, 2002.