Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 04-008

Impact of AS Hierarchy on Multihoming Performance: A Stub Network Perspective

Sanghwan Lee, Zhi-li Zhang, and Srihari Nelakuditi

February 20, 2004

Impact of AS Hierarchy on Multihoming Performance: A Stub Network Perspective*

Sanghwan Lee, Zhi-Li Zhang
Computer Science Department
University of Minnesota
200 Union Street
Minneapolis, 55455 USA
{sanghwan, zhzhang}@cs.umn.edu

Srihari Nelakuditi
Dept of Computer Science & Engg
University of South Carolina
Columbia, SC 29201
srihari@cse.sc.edu

ABSTRACT

Multi-homing, namely, connecting to more than one Internet Service Provider (ISP) for global Internet reachability, is a common practice among many (especially large) customer (or stub) networks. Although the purpose of multi-homing is primarily for enhanced reliability, it has also increasingly been used for load balancing and other performance benefits. This paper is motivated by the following major question: in a multi-homed stub network, is there any significant benefit in carefully selecting one of the several available ISPs to optimise latency (as measured by round trip time, RTT) to various destinations? To answer this question, we carry out a measurement-based study to compare and analyze performance differences in using two different providers in a multi-homed stub network to reach a large number of randomly selected destinations. Our study reveals that there are often performance benefits in selecting the best provider to optimise network latency. Furthermore, for a large fraction of the network prefixes, the RTT differences between the two providers fall into a dominant range. This phenomenon can be attributed to the effect of the AS hierarchy on AS paths: the AS hierarchy often causes the AS paths via the different providers to merge at the core of the Internet, resulting in shared common segments to many network prefixes and ASes. Consequently there is strong correlation among RTTs to many destination networks. Our findings provide some useful insights as to how to perform intelligent provider selection using BGP in a multi-homed stub network.

1. INTRODUCTION

The Internet is a collection of separately administered network domains, in routing jargon, *autonomous systems* (ASes). Some ASes are ISPs which provide so-called transit service, namely, carrying traffic from one part of the Internet across their networks to another part; while the majority of them are networks owned by corporations, universities, government agencies, and other organizations that buy transit service from ISPs for global Internet connectivity. These latter ASes are often referred as *customer* networks (or *stub* networks since Internet traffic only originate or sink at these networks). The Border Gateway Protocol (BGP) is the de facto standard inter-domain routing protocol used among ASes to exchange routing information for global reachability. It is a path vector based protocol designed with the objective of allowing ASes to apply their own routing policies in selecting and propagating routes. Because of the various business relations formed and routing policies used by ASes, it is now well known that the Internet AS topology reveals certain inherent hierarchical structure. For example, in [11] the five-level AS hierarchy is proposed by which ASes are classified into *Dense Core, Transit Core, Outer Core, Regional ISPs* and *Customer Networks*.

Many customer networks, especially large ones, are often connected to multiple ISPs. This practice is referred to as *multi-homing*. The primary objective of multi-homing is to provide enhanced reliability in the event of failures in a provider network. Since a multi-homed customer network must pay for the connections to multiple ISPs, it is desirable to also exploit multi-homing for performance optimization such as load balancing. Network latency is another important performance metric that can be combined with load balancing. For example, a multi-homed customer can choose to route traffic via one of the providers to those destination networks that are closer to the provider and thus afford shorter network latency, while at the same time properly balancing the amount of traffic routed through each provider. Such performance optimization via "intelligent" provider selection is particularly beneficial to multi-homed customer networks with large web hosting or data center facilities, where reduced network latency can significantly improve the overall performance of application servers in such web hosting and data center facilities. To enable "intelligent" provider selection, a multi-homed customer network can apply certain routing policies in the BGP path selection, e.g., by appropriately setting the `LOCAL PREFERENCE` attribute (see section 2 and [7]).

This paper is motivated by the following questions: 1) in a multi-homed stub network, are there any significant benefits in carefully selecting one of the several available ISPs to optimize latency (as measured by round trip time, RTT)

*Sumitted to Sigmetrics 2004

to various destinations? 2) if the answer to 1) is affirmative, what are the possible factors that contribute to such benefits? and 3) how can such benefits be effectively employed? In particular, at what granularity can we apply BGP routing policies to take advantage of such benefits? Towards answering these questions, we set up a small measurement infrastructure in a multi-homed stub network with two commercial providers and conducted a series of experiments to collect RTT measurement data to a large number of destination networks (network prefixes) via each of the two providers. Using the datasets we collected, we compared and analyzed the difference in the RTT performance through these two providers, and investigated what are the potential factors that contribute to the performance difference by correlating the RTT measurement with traceroute data and BGP routing information. Our major findings and contributions are summarized below.

- There are indeed *considerable performance differences* in terms of RTTs in choosing different providers to route traffic from a stub network to other destination networks. More interestingly, we find that there is a *dominant* range into which the most of differences in the RTT performance between the two providers fall. For example, our measurement data show that for a large fraction of network prefixes, one provider affords a better RTT performance in the range of 15 ms to 25 ms over the other provider. This RTT performance gain is independent of where the destination networks are located in the Internet AS hierarchy or how far they are from the stub network in which the measurement was conducted. The AS Path length, used by BGP as the second path selection criterion after `LOCAL PREFERENCE` attribute, in fact often does not yield a better path in terms of network latency, as is evident in the fact that for a sizeable portion of network prefixes belonging to one provider, using the other provider actually results in shorter RTT values.

- The above observation is not only true at the prefix level, but also at the AS level. Namely, for a large portion of ASes, choosing one provider over the other provider to route traffic to *all* the prefixes in those ASes is more *preferable*, as it yields a considerable overall gain in delay performance. This is particularly true for ASes that are small or are at the lower levels of the AS hierarchy (e.g., customer networks). Whereas, for large ASes, in particular, in the Dense Core, in general no one provider consistently outperforms the other provider.

- Furthermore, our analysis reveals that the AS hierarchy has a major impact on the RTT performance, in that the AS hierarchy often causes the AS paths via the different providers to merge at the core of the Internet – the Dense or Transit Core, resulting in shared common segments to many network prefixes and ASes. Due to these shared AS path segments, the RTTs among different network prefixes are often correlated, and the difference in the RTT performance via different providers are determined to a large extent by the RTTs from the stub network to the Dense Core or Transit Core of the Internet AS hierarchy. We provide strong evidences that this phenomenon of shared AS path segments due to the AS hierarchy is very prevalent by performing *virtual* stub network analysis using the routeview BGP data [3]: we identify many ISPs that have PoP presence in several major cities in the world, and investigate how the charcteristics are similar to our stub network.

Our findings have several important implications in intelligent provider selection using BGP. First, it shows that it is worthwhile to carefully select the best provider to reach certain destination networks to optimize network latency. Second, such provider selection can often be done at the AS level, instead of prefix level, making the task of configuring appropriate routing policies for provider selection easier and less tedious. More importantly, our results also suggest that instead of performing RTT measurement to every single network prefix, we can exploit the correlation in RTTs to network prefixes with shared AS path segments to reduce the amount of measurement and monitoring traffic, thereby making intelligent provider selection for delay performance optimization more scalable.

Lastly, we remark that although our findings are based on the measurement data from a single stub network, we believe that they are not unique to this stub network, and the *general* observations should hold also in other stub networks. In particular, our major finding regarding the impact of the AS hierarchy on the RTT performance via different providers should have broad applicability, as is corroborated by our analysis of BGP data from the routeview project [3]. Clearly more extensive measurement using more multi-homed stub networks is needed to further verify and confirm our findings. Nonetheless, this paper is the *first* study that uses extensive measurement data collected through experiments in a *real* multi-homed stub network to explore the performance gains in provider selection (in terms of network latency) and to analyze the potential factors contributing to such performance gains. Our discovery that the AS hierarchy plays an important role in the RTT performance differences sheds *new* insight on how *effective* and yet *scalable* provider selection can be employed in network latency optimization.

The remainder of this paper is organized as follows. Section 2 provides some background on BGP and AS hierarchy, and describes our measurement set up and experiment methodology. In section 3, we compare and study the differences in the RTT performance via the two providers. In section 4 we correlate the RTT and traceroute measurement data to analyze the impact of the AS hierarchy on the RTT performance. Section 5 discusses the implications of our findings on the BGP path selection. Section 6 concludes the paper with a brief discussion on the future work.

1.1 Related Work

As multi-homing is increasingly adopted by customer networks for enhanced reliability and other performance benefits, companies such as [2, 1] are offering commercial products to exploit dynamic provider selection for improved performance. From what we have learned from such products (e.g., via the company websites), they typically rely on continuous monitoring and active measurement of user network performance. Typically of most such commercial products, no scientific studies have been published demonstrating their effectiveness.

The first academic research that provides a systematic analysis of the potential performance benefits of multi-homing

is the excellent recent study in [4], where extensive measurement data collected from web hosting and data center facilities are used. Using the measurement data, the authors investigated the performance benefits of multi-homing by comparing the performance (e.g., response time) of web servers located *in the same city* but connected to different ISPs, and treated them *as if they were multi-homed*. Their study shows that there are significant performance benefits in *dynamic* provider selection based on previous delay measurements. The performance gains in general increase as the number of providers grows up to 4. However, after 4 providers, the gains become insignificant. Our work is similar in the spirit to [4], but differ in several important ways: 1) our measurement experiments are conducted in a real multi-homed network, 2) the multi-homed network is a stub network, as opposed to co-locating facilities near/within the core of the Internet; and 3) our focus is on understanding the factors such as the AS hierarchy that contribute to the different performance of upstream providers, instead of on the number of providers used. Hence our study complements and further advances the study of multi-homing pioneered by [4].

In another piece of related work, the study in [5] uses active measurement through different gateways to model and predict the delay performance to avoid degraded services via dynamic provider selection. Their study is based entirely on active probing. BGP route data is not used to correlate and understand delay performance. In another recent study, [10] shows the paths selected by BGP are often *not* optimal in terms of network latency, due to factors such as routing policies (such as "early exit") and preference of shortest AS path. The findings in this study not only provide insight into our work, but are also further confirmed by our study.

2. BACKGROUND, MEASUREMENT SETUP AND EXPERIMENTS

In this section we first provide some necessary background on Internet AS hierarchy and BGP. We then describe our measurement set-up, the experiments and data we conducted and collected for our study.

2.1 BGP and Internet AS Hierarchy

The Internet consists of more than 14000 network domains, or ASes (autonomous systems). BGP is the de facto inter-domain protocol used among ASes to exchange routing information for global Internet connectivity [7]. BGP uses a basic path vector protocol to announce route updates – a list of ASes traversed, called AS path is included (among many other attributes) in the route updates as they propagate among ASes; the primary objective of BGP is to enable ASes to apply *policy control* in route selection, filtering and propagation.

Because of the different business relations (e.g., customer-provider, peering [8]) formed among ASes and the resulting policies they use to filter, select and propagate BGP route updates, it is now well known that the (logical) Internet AS topology reveals a hierarchical structure, with a few ASes (so-called tier-1 ISPs) constituting the core of the Internet, where most of the Internet traffic traverse. In this study we adopt the classification of the Internet AS hierarchy proposed in [11], which categorize ASes into five levels: *Dense Core* (level 1) consists of top tier-1 ISPs that form a full mesh; *Transit Core* (level 2) consists mostly of large ISPs that have rich connectivity to the Dense Core but are full-meshed with them (i.e., they are likely to have customer-provider relations with some of the top tier-1 ISPs in the dense core; *Outer Core* are mostly medium-sized ISPs which are not as richly connected as the ISPs in the transit core and are further away from the core of the Internet; *Regional ISPs* are mostly small ISPs with only regional presence; and *Customer Networks* are ASes that do not *transit* traffic for other ASes, i.e., they are the leaf nodes in the AS graph (thus also called *stub* networks). The majority of ASes are customer (or stub) networks. Note that the classification methodology used in [11] is based on BGP data collected at multiple vantage points. Since BGP routes change over time, the classification of an AS may also vary over time.

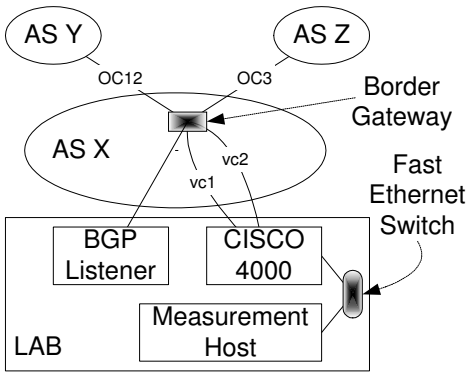

Figure 1: Multihomed Measurement Environment

When an AS is connected to multiple ASes, it will likely receive multiple route announcements to a given destination network prefix, of which the AS selects the "best" route based on policy and other considerations. The standard BGP best route selection uses the following criteria [7]: first the route with the highest `Local Preference` attribute value is selected; If the `Local Preference` values are the same, the route with shortest `AS Path` is selected; if the `AS Path` length is the same, then the MED (Multi-Exit Discriminator) attribute value, IGP (i.e., intra-domain routing) metric, the `Origin` attribute and router id are considered in the order they are listed (see [7] for more details). An AS typically influences the route selection process by manipulating the `Local Preference` attribute. Best route selection decisions are generally made based on routing policies (which are often driven by monetary consideration and business relations). Performance is usually a secondary concern. This is particularly true for ASes that are ISPs. However, as mentioned in the introduction, in multi-homed customer networks performance can be an important consideration in deciding which route to use to reach certain destinations. To effectively take performance such as network latency into account when making route selection decisions, we need to understand the major factors that affect the network performance. Such knowledge can help us make intelligent route selection decisions, e.g., by setting appropriate `Local Preference` attribute value to control which provider to use to reach various destinations.

2.2 Measurement Setup

For our study we set up a small measurement infrastruc-

ture in a multi-homed campus network (see Fig. 1). The campus network is connected to two commercial ISPs. For anonymity, the campus network is referred to as AS X, and the two commercial ISPs are referred to as AS Y and AS Z, respectively. Both AS Y and Z belong to the Transit Core. Two virtual circuits (through a dedicated fiber connection over an ATM network) are established between a Cisco 4000 router in our lab to the border gateway router of the campus network. A measurement host (a Linux PC) is connected (via a Fast Ethernet switch) to the Cisco 4000 router. The measurement host is assigned two special IP addresses: packets with one of these IP addresses as the source address will be sent through one of the two VCs. The border gateway router and the Cisco 4000 router are configured such that by choosing one of the two VCs to send packets, the measurement host can control which provider, AS Y or AS Z, to use for sending packets to any given destination on the Internet. Another host (the "BGP listerner") in the lab establishes a BGP session with the border gateway router, and receives all the BGP route updates (the "best" route through one of the two providers selected by the border gateway router to every destination prefix in the Internet). We note here that the connection from AS X to AS Y is an OC12 circuit, and from AS X to AS Z is an OC 3. Hence not surprisingly, most traffic originating from AS X are sent through AS Y, as can be verified by the BGP routes received by the BGP listener.

We wrote a simple measurement program to measure the round trip time (RTT) from the measurement host to a given destination address via each of the two providers, AS Y or AS Z. The measurement program launches two `ping` processes concurrently: one `ping` process sends a predefined number of `ping` probe packets within a constant interval via AS Y, and the other via AS Z. Fig.2 depicts how the `ping` probes are sent: the interval a is the time between the probes sent by the two `ping` processes, while the interval b is the time between two consecutive probes sent by the same `ping` process. After the RTTs of all `ping` probes to one destination are recorded, a new set of two `ping` processes are launched for another destination. In section 2.3 we will describe in more detail how the destinations are chosen, and how the RTT measurement experiments are conducted. In addition to the `ping` measurement, we also run *traceroute* measurement to collect the information about the (router-level) paths the `ping` probes take to reach a destination through each of the two providers. From the router-level paths, we also obtain the AS-level path information by mapping the IP addresses to their ASes using BGP data and AS mapping tool [9]. Note that the BGP listener only gives us the AS path information to a destination via one of the two providers (mostly via AS Y), but not both. The traceroute data provides us with the AS path information the `ping` probes take via the other provider. The path information (both at the router level and AS level) proves crucial to our uncovering the impact of AS hierarchy on multi-homing RTT performance in section 4.

2.3 Experiments and Data Processing

We conducted `ping` measurement experiments in three stages. The purpose of the first stage experiment is to discover "live" IP addresses and collect initial RTT measurement data. Using BGP routing data collected on Aug. 15, 2003, we choose the set of all the network prefixes, and

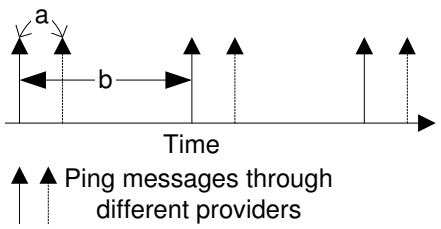

Figure 2: Interval and Sequence of Ping Messages

	stage 1	stage 2	stage 3
Start Date	Aug.15 03	Sep.30 03	Oct.15 03
End Date	Sep.19 03	Oct.5 03	Oct.29 03
Interval a,b	0,4.5(s)	1(s),2(s)	1(s),2(s)
Order	Y,Z	Z,Y	Z,Y
# messages	3	3	10
# IP tried	246,932	65,631	65,631
# IP responded	65,631	60,145	57,721
Final Dataset	36,219 prefixes		

Table 1: Measurement Parameters

from each prefix, we randomly select two IP addresses (as in [9]). The resulting target IP address set contains a total of 246,932 IP addresses. The first-stage experiment was done from Aug. 15, 2003 to Sep. 19, 2003: for each destination in the target IP address set, two `ping` probes were sent back-to-back, one via AS Y, and the other via AS Z; three `ping` probes are sent every 4.5 seconds via each provider (i.e., interval $a = 0$ and interval $b = 4.5$). Out of the 246,932 IP addresses tried, 65,631 responded.

In the second stage (from Sep. 30 to Oct. 5, 2003), we repeated the same experiment using only those 65,631 IP addresses that responded in the first stage experiment as the target IP address set. We also changed the order of the providers we used (first via AS Z, then AS Y) to ping a host, and the intervals a and b used are 1 second and 2 seconds, respectively. The purpose of the second stage experiment is to verify the RTT data we collected, and to check whether they are sensitive to the order and the frequency in which the `ping` probes are sent. The third stage experiment were done from Oct 15 to Oct 29, 2003. This time the purpose is simply to test consistency of the RTT data and check to see whether they change over time. Hence we sent 10 `ping` probes to each destination via each provider. The statistics of the three experiments are summarized in Table 1.

The three-stage experiments yield two sets of RTT measurement data, one for each provider. Each data set contains *up to* 16 RTT `ping` measurements to each of the 57,721 IP addresses responded in the third stage. (We discarded the earlier RTT measurements to those IP addresses that did not respond in the third stage.) Using these two datasets, we then performed the following data analysis and filtering to verify data reliability and remove "outliers". Our objective is to obtain a pair of *representative* RTT values, one for each provider, to reach a given *network prefix*, so that we can analyze the performance benefit in "fine-tuning" BGP route selection to optimize delay performance. Each RTT value should ideally reflect the minimum network latency (i.e., propagation delay) to reach the network prefix via each

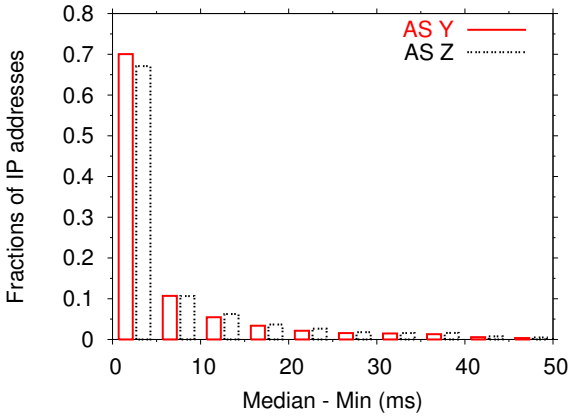

Figure 3: Distribution of median - min of 16 RTTs

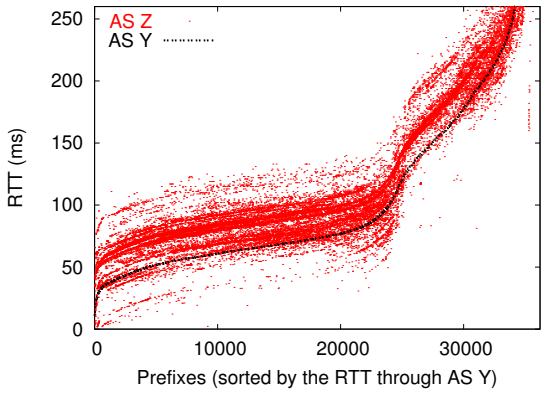

Figure 4: RTTs to the prefixes

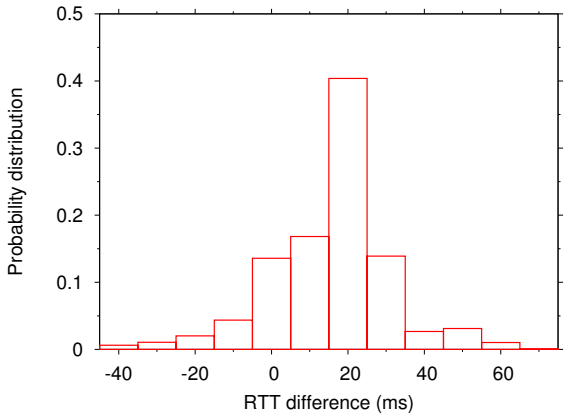

Figure 5: Distribution of RTT Differences:(Median, Mean, STDev) = (19.5ms, 16.1ms, 19.0ms)

provider.

In each dataset, we first discard those IP addresses with fewer than 16 `ping` responses, which could indicate that the network was congested (or the host was busy or unreachable) at some point. For each IP address with 16 RTT measurement data, we compute the minimum and median of the RTTs and use their difference, *median - min*, as a measure of degree of variation. Fig. 3 shows the "distribution" of the resulting *median-min* values for both datasets, where the *median-min* values were grouped into 5ms bins and the fraction of IP addresses fall into each bin is shown. We see that the RTT data that we collected over a time span of two months are fairly consistent, as an overwhelming fraction of them have small *median-min* values. Since the bin that contains the 95th percentile is [35ms,40ms), we consider the *median-min* value larger than 40 ms as "outliers" and remove the corresponding IP addresses from both data sets. For the remaining IP addresses in each dataset, we use the minimum of the 16 RTT values as the *representative* RTT value. Finally, recall that for each *network prefix*, we selected two IP addresses for the `ping` measurements. If a network prefix has two IP addresses in the datasets, we then use the smaller of the two RTTs as the *representative* RTT to the said network prefix. The final dataset consists of the RTT value pairs of *36,219* network prefixes. For each network prefix p in the final dataset, we use *rtt(p, Y)* to denote the (representative) RTT from our measurement host to p via the provider Y, and *rtt(p, Z)* the RTT via the provider Z.

3. PROVIDER DELAY PERFORMANCE COMPARISON

In this section we study the difference in the RTT delay performance to reach a large set of destination network prefixes via the two providers. We first analyze whether factors such as their distance to our measurement hosts (in terms of RTT), where the network prefixes are located in the AS hierarchy as well as the AS path length to reach these prefixes play a role in the decision in provider selection for delay performance optimization. We then investigate the question whether one provider always performs better for all the prefixes in an AS than the other provider. Such knowledge can help us understand the granularity (prefix or AS level) at which we could select the provider to optimize delay performance.

3.1 Prefix-Level Performance Comparison

Fig. 4 shows the RTTs from the measurement host to the 36,219 network prefixes in the final dataset via each of the two providers, where the x-axis is numbered by the prefixes ordered based on their RTTs to the provider Y. This is why the RTTs via AS Y form a continuous curve in the figure, whereas the RTTs via AS Z are points scattered around the curve. Clearly, to most network prefixes going through AS Y in general yields shorter RTT than through AS Z. It is also notable that there seems to exist a visible *band* about 20 ms or so wide above the curve (RTTs via AS Y) formed by RTTs via AS Z. This seems to indicate that for a large portion of the network prefixes, the delay difference between the two providers falls within 20 ms or so. This is confirmed by the RTT difference distribution.

For each prefix p in the dataset, we compute the difference in the RTT via the provider Z and via the provider Y, i.e., $rtt(p, Z) - rtt(p, Y)$. This RTT difference reflects the difference in using the provider Z vs. provider Y to reach the network prefix p. The distribution of the RTT differences is shown in Fig. 5, where the prefixes are grouped into bins of 10 ms width, such as (-15,-5],(-5,5), [5,15),[15,25), based on their RTT differences, and fraction of prefixes in each bin is shown. It is clear that more than 80% of the

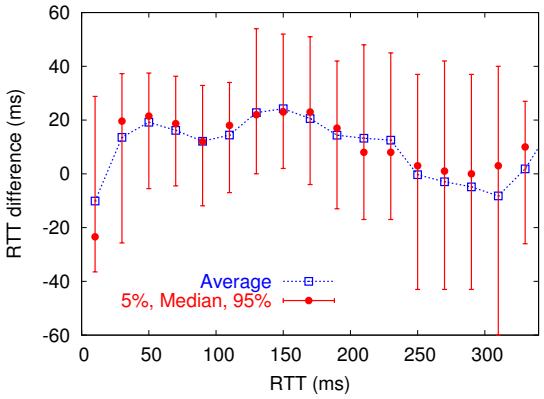

Figure 6: Distribution of RTT Differences over "Distance" to Prefixes

prefixes are in (-5ms, 35ms) range and the largest fraction is in the range of [15ms, 25ms), which consists of more than 40% of the RTT differences, which contributes to the band we observed in Fig.4. We conclude that although the RTT differences between the two providers are not huge, they are large enough that preferring one provider over the other will provide some latency benefit.

We now investigate whether the RTT difference via the two providers is correlated to the "distance" of a destination network to the stub network. To measure the "distance", we uses the minimum of the two RTTs[1], i.e., $\min\{rtt(p,Y), rtt(p,Z)\}$, and then group the prefixes into 20ms-bins accordingly: [0ms, 20ms), [20ms,40ms), [40ms, 60ms), etc. Fig.6 shows the average, median, 5th, and 95th percentile of the RTT differences for the prefixes in each 20ms-bin. One might expect that when the distance of the prefix is larger, the RTT difference between the two providers may also be larger, that there would be more benefit to use one provider over the other. This in general is not the case in our dataset. In the range of [20ms, 200ms), the difference distribution does not differ very much: the averages and medians oscillate around 20ms. When the "distance" is larger than 200ms, the average differences fall between 20ms. This suggests that the distance from a stub network to a destination network does not in general favor one provider over the other.

Using the AS hierarchy classification scheme in [11] and discussed in section 2, we group the prefixes based on the categories their originating ASes belong to: Dense Core, Transit Core, Outer Core, Regional ISPs, and Customer Networks. Similar to Fig. 5, we plot the distribution of the RTT differences for each category. In Fig.7, the distributions for Dense Core, Transit Core, Outer Core, and Customer Network (for readability, the figure does not include that for Regional ISPs, but it looks similar to the other distributions). We find that across all categories, the the distribution fairly similar to each other and that of Fig. 5. Hence the level/position in the AS hierarchy a prefix's originating AS belongs to does *not* seem to have a strong impact on the RTT performance difference in using one provider or the other.

[1]In [10], it is shown that RTTs are predominantly determined by geographically locations, with some "inflation" (generally less 20%) caused by routing policies and other factors.

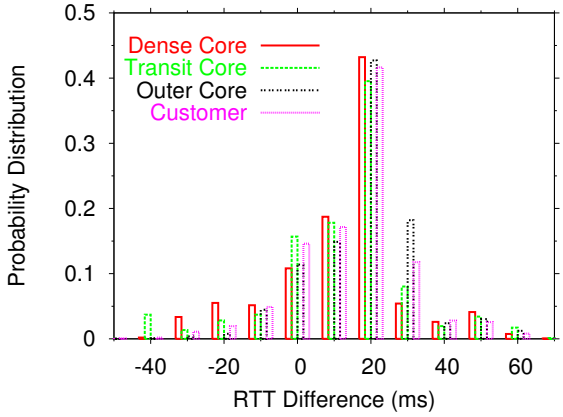

Figure 7: Distribution of RTT differences over AS hierarchy positions

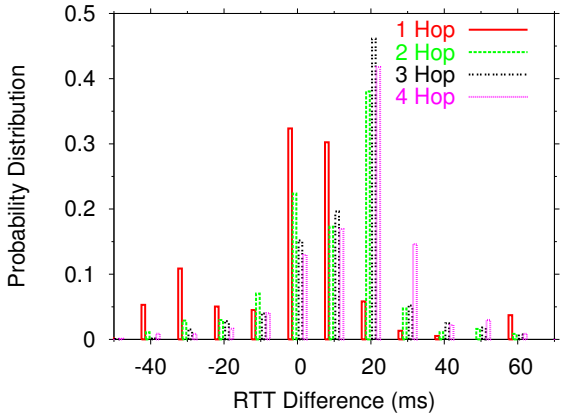

Figure 8: Distribution of RTT differences over AS path length

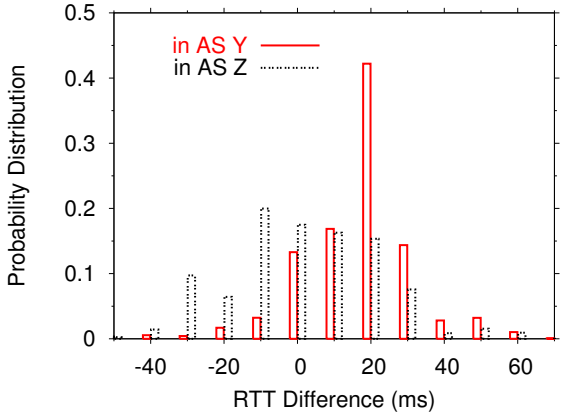

Figure 9: Distribution of RTT difference to 1 Hop destinations

Finally, we turn to the relationship between the RTT performance difference and the length of the AS path used by the (default) provider (i.e., the one chosen by the border gateway router in the stub network) to reach a destination network. The AS path information is extracted from the BGP routing table that collected by our BGP listener. The prefixes are then grouped by their AS path lengths. Fig.8 shows the distributions of the RTT differences over the AS path lengths. Overall, except for the *1 Hop* prefix group, all other groups have fairly similarly distributions, with the largest fraction concentrated in the RTT difference range [15ms, 25ms]. For the *1 Hop* prefix group, the largest fraction is in the range [-5ms,5ms]. Note that the prefixes with 1 hop AS path belong to either provider Y or Z. Intuitively one would expect that selecting the provider from which a prefix originates should yield better performance. Our result shows that for more than 30% of the prefixes belonging to either the provider Y or Z, using the other provider does not have a strong negative impact on the RTT delay performance. Fig.9 shows the distribution of the RTT differences of the prefixes only in AS Y and AS Z. More interestingly, Fig.9 shows that in fact for a sizeable fraction (larger than 20%) of the prefixes belonging to the provider Z, the RTT difference is larger than 15ms, which implies that going through the provider Y to reach those destination networks in AS Z is at least 15ms faster than going through the provider Z. We suspect that the reason for such phenomenon is due to the fact that both AS Y and AS Z are fairly large ISPs covering a large geographical area. Hence for each of them, a sizeable fraction of its networks are closer to the PoPs (Points-of-Presence) of the other AS, thus resulting in shorter network latency by going through the other provider. These results further confirm the finding in [10] that the shorter AS path may not guarantee a shorter network latency.

We conclude this section by summarizing our findings. Our results show that in terms of network latency (as measured by RTTs) there is in general considerable performance difference in using one provider over the other provider. For a large fraction of network prefixes, one provider outperforms the other with a RTT difference in the range of 15 to 25 ms. This performance difference does not strongly depend on the distance to the destination networks, and where in the AS hierarchy their ASes reside. Except for the prefixes belonging to the two providers, the AS path length does not appear to play a role either. In section 4 we will perform a more in-depth analysis of the correlation between RTT performance difference and AS paths to gain more insight into these findings.

3.2 AS-Level Performance Comparison

So far we have analyzed the RTT performance difference in using one provider over the other to reach a destination network. In this section we study the problem whether one provider yields better delay performance over the other for *all* the prefixes in a given AS. We first introduce some notation and define a performance metric.

For each prefix p, we define an indicator function, $I(p, Y)$, as follows. For a fixed $\delta > 0$,

$$I(p,Y) = \begin{cases} 1 & \text{if } rtt(p,Y) - rtt(p,Z) > \delta,\ u \neq v \in \{X,Y\} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Namely, $I(p, Y)$ represents whether going through the provider Y will yield a delay gain of at least δ ms over the provider Z, thus it should be the *preferred* provider to reach p. The indicator function, $I(p, Z)$ is similarly defined.

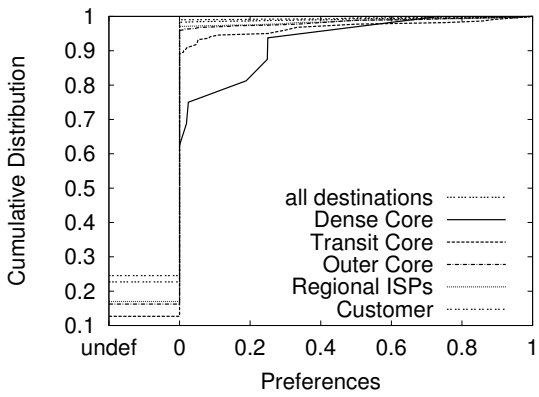

Figure 10: Preference over AS Hierarchy ($\delta = 10ms$)

For a given AS S, let $P(S)$ denote the set of prefixes in S. For $U \in \{Y, Z\}$, let $I(S,U) = \sum_{p \in P(S)} I(p,U)$. Hence $I(S,Y)$ is the number of prefixes in S that the provider Y is preferred over the provider Z, and $I(S,Y)$ is the other way round. We now introduce a performance metric, called *preference meter* and denoted by $pref(S)$, to measure the "degree" to which one provider is more preferred over the other for AS S. It is defined as follows.

$$pref(S) = \begin{cases} \frac{\min\{I(S,Y),I(S,Z)\}}{\max\{I(S,Y),I(S,Z)\}} & \text{if } \max\{I(S,Y),I(S,Z)\} \neq 0 \\ undefined & \text{otherwise} \end{cases} \quad (2)$$

Note that $pref(S) = 0$ means that one provider is always preferred than the other provider, whereas $pref(S) = 1$ means that a provider performs better for some prefixes and the other provider performs better for the other prefixes. That $pref(S)$ is *undefined* means that both providers have roughly equal performance for all prefixes in AS S, thus there is no preference between the two. Using this performance metric, we now investigate whether there is considerable performance difference at the AS level by selecting one provider over the other, and whether where in the AS hierarchy an AS belongs to, its size (in terms of number of prefixes), and the AS path length to reach it play a role in preferring one provider over the other.

Using $\delta = 10$ ms, Fig.10, 11, and 12 show the cumulative distributions of the *preference meters* for various groups of ASes categorized using three different criteria: in Fig.10, the ASes are grouped based on the level of the AS hierarchy they belong to; in Fig.11, they are grouped based on their size (number of prefixes they contain); and in Fig.12, the ASes are grouped based on the (default) AS path length. In all the cases, the preference meter for around 20 % of the ASes is undefined, meaning that there is no significant performance difference (less than 10 ms) in using either of the providers. For a majority of the remaining ASes, the preference meter is 0. Hence for these ASes, one provider yields significantly better performance than the other. Comparing ASes in the Dense Core vs. in the other levels of the AS hierarchy, the fraction of ASes that one provider yields better performance than the provider is much lower. The same

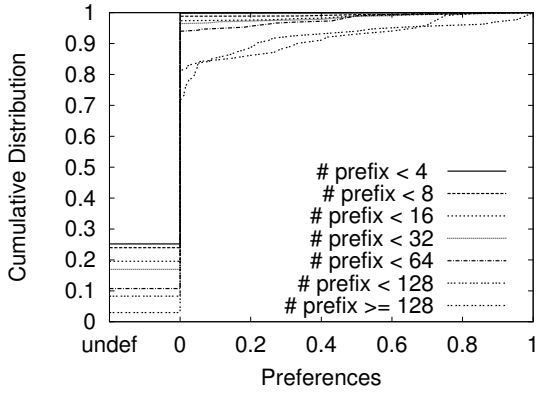

Figure 11: Preference over AS size ($\delta = 10ms$)

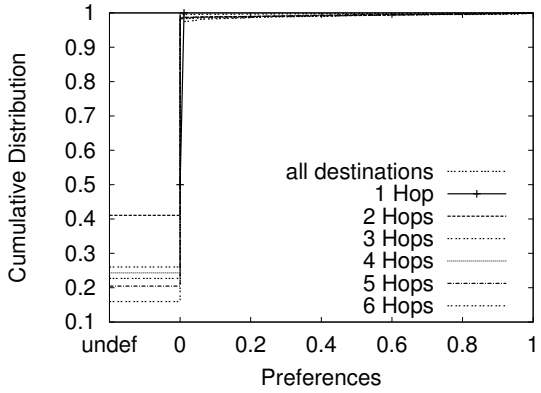

Figure 12: Preference over AS Path Length ($\delta = 10ms$)

applies to the ASes with large size (more than 64 prefixes). This is not too surprising as most ASes in the Dense Core are large ASes, and large ASes tend to cover a large geographical span. As a result, for certain prefixes within these ASes one provider is likely to have PoPs closer to these prefixes than those of the other provider. In contrast, for ASes in Customer Networks (and also Regional ISPs) or with small sizes, one provider tends to yield better delay performance than the other. As the majority of ASes belong to the group of Customer Networks, the overall distribution of preference meters looks similar to that of Customer Networks, hence a large majority of ASes have a preference meter of 0. Lastly from Fig.12, we see that except for 1 Hop ASes (which consists of only AS Y and AS Z, the two providers), AS path length does not have any impact on the distribution of preference meter. This implies that selecting providers based on AS path length (the second criterion after LOCAL PREFERENCE attribute used in BGP path selection) does *not* have any bearing on the delay performance.

4. THE IMPACT OF AS HIERARCHY

In the previous section, we raised several interesting questions: Why does one provider perform better than the other provider for all the prefixes of an AS? Why is the difference in RTTs via the two providers to a destination uncorrelated to their absolute RTT values? Why is there a dominant range for these RTT differences across a large set of destinations? The answers to these questions would provide guidance in choosing a set of service providers and in assigning local preferences to these providers for selection of best paths from a multihomed stub network. In this section, we analyze the path information gathered through our traceroute experiments and show that the AS hierarchy holds the key to answering all the above questions. We observe that many AS paths through different providers from a multi-homed stub network share common segments and the RTT from a stub network to many destinations is often determined by how it reaches the core Internet. We further analyze the AS path information collected from Routeviews [3] and show that these observations based on our measurements from a single multi-homed stub network are applicable in general for any stub network.

It is well known that the Internet AS topology reveals certain inherent hierarchical structure because of the various business relations formed and routing policies used by ASes. At the top of AS hierarchy of the Internet, there is the dense core in which all the ASes have full-mesh peerings. Since majority of the ASes are the customers of the ASes in the dense core, it is very likely that the many AS paths contain the ASes in the dense core. As a consequence, *two paths from a stub network via two providers to a destination are likely to merge at the dense core.* Fig.13 illustrates an AS topology where two paths through different providers from a source S to a destination D merge at an AS in the dense core. We carefully analyze in the following where in the hierarchy two paths via different providers to a destination merge as it helps provide answers to the above questions.

4.1 Analysis of path correlation

Let *merging point* be the node at which the two paths merge. Fig.14 shows two paths from a source S via providers X and Y to a destination D that merge at node M. This merging point could be the destination D itself if the paths

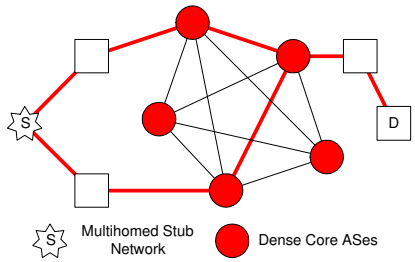

Figure 13: Merging At Dense Core

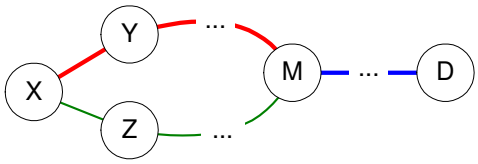

Figure 14: Merging AS paths

via Y and Z are completely disjoint. Let the two paths from the source to the merging point be the *branch path pair*. Since in most cases, there is one-to-one correspondence between the merging point and the branch path pair, we use both terms interchangeably. In the Fig.14, the pair of paths XY..M and XZ..M is the *branch path pair* for the two paths from X to D. These paths can be AS level, prefix level, or IP level paths and thus the merging point can be an AS, a prefix or an IP address depending on the context[2]. We now make use of this notion of merging point and branch path pair in analyzing the path information collected through traceroute experiments from X via Y and Z to several destinations. Since some routers did not respond to traceroute queries we collected path information to 31,188 prefixes out of the 65,631 prefixes.

The path selection under BGP is performed at the granularity of the prefixes. It is not necessary that all the prefixes of an AS share the same path. However, from the viewpoint

[2]Since a router can have different IP addresses, the actual paths may merge earlier than the merging IP. However, the difference will be at most 1 hop.

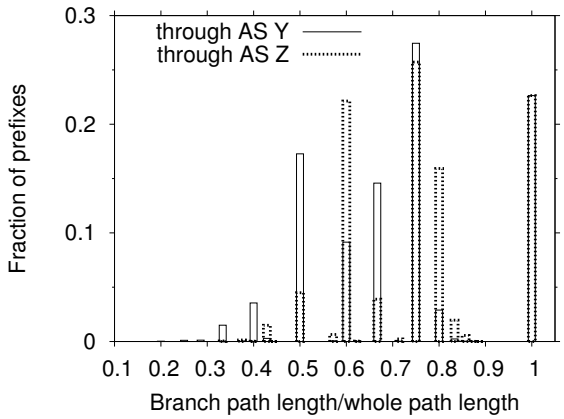

Figure 15: Ratio of the branch path length over the whole path length

Position	Fraction
Dense Core	0.766
Transit Core	0.088
Outer Core	0.0470
Regional ISP	0.044
Customer	0.056

Table 2: Fraction of prefixes merging at each hierarchy position

of a stub network X, if the performance of one provider Y is better than the other provider Z for all the prefixes of an AS D, then X can select the best paths by simply assigning a higher local preference to Y for all prefixes of D instead of configuring it for each prefix. The results from the previous section have indeed shown that one provider yields better delay performance over the other provider for all the prefixes in an AS. Here we would like to better understand why this is the case. For that purpose, we consider the ratio of the lengths at AS level of the branch path and the whole path to a destination. If this ratio is 1, it implies that the paths via two providers are disjoint and merge only at the destination AS. Otherwise they merge prior to the destination AS. Fig.15 shows the distribution of this ratio. It can be seen that only around 20% of the prefixes have the ratio 1, i.e., their paths merge at the destination. For more than 75% of the prefixes their paths merge along the way. Therefore the path pairs to prefixes of an AS are likely to merge prior to the AS and the merging points can be same for all the prefixes. This explains why the RTT differences to all the prefixes of an AS are similar and one provider has better delay performance than the other for the whole AS.

The delay measurements presented in the previous section have also indicated that the performance gain afforded by a provider to a destination AS is independent of the distance to that AS from our stub network. We now investigate the causes behind this phenomenon. As explained before, if two provider paths merge along the way to a destination, their RTT difference is dictated by the branch paths to the merging point. Fig.15 shows that large fraction of the ratios lie between 0.5 to 0.8, i.e., path pairs are likely merge between half-way and three-quarter-way to the destination AS. To figure out where exactly in the hierarchy two paths meet, we mapped the merging points to their position in the hierarchy. The fraction of paths that merge at each hierarchy position are given in Table 2. It can be seen that for more than 75% of the prefixes, path pairs merge at the dense core. So it is likely that path pairs to several destinations merge at the same node in the dense core resulting in their RTT difference being similar irrespective of their distance from the merging point.

To corroborate the above intuition we grouped all the destinations according to their merging points, i.e., if two destinations share the same merging point they are included in the same group. Fig.16 displays the range of absolute RTTs for destinations within each group. Only the groups with more than 20 members are shown in the graph for the purpose of clarity. It is clear that destinations with wide range of RTTs may have the same merging point. We then considered the RTT differences to destinations within each group. Let DRTT of a destination be the difference in RTT between two provider paths to that destination. We declare a group

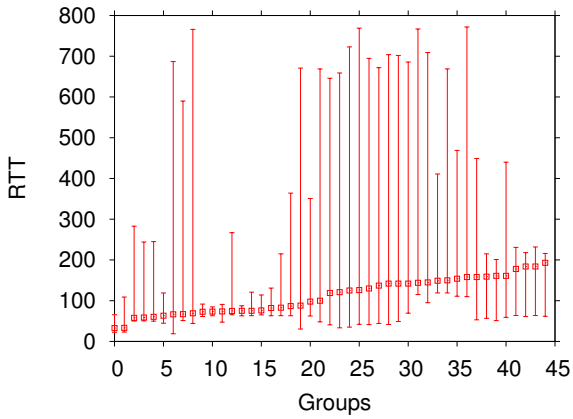

Figure 16: Min, Median, and Max of RTTs in each group categorized by IP level merging points, groups are sorted by their Median

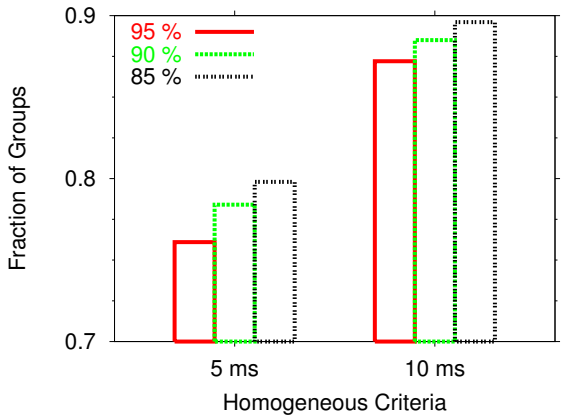

Figure 17: Fraction of homogeneous groups based on merging point

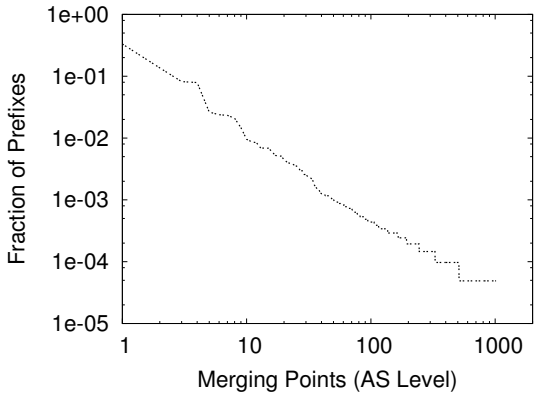

Figure 18: Merging Point Sizes (Traceroute)

as *homogeneous*, if more than a certain *threshold* fraction of the destinations have their DRTTs within an *error margin*. Fig. 17 presents the fraction of homogeneous groups for threshold values of 80%, 90%, 95%, and error margins of $5ms, 10ms$. Across all the cases, at least 75% of groups have similar DRTTs. It is clear from contrasting figures Fig. 16 and Fig. 17 that destinations sharing a merging point have similar DRTT even though they have quite different RTTs. Therefore we can conclude that due to the likelihood of merging of various paths inside the core, the delay gain attained through a provider is independent of the distance to a destination.

We now turn our attention to another interesting finding from our measurements that there is a dominant range of DRTTs across all destinations. This can be explained based on the above observations and AS hierarchy structure as follows. We have seen that the DRTTs of destinations that share the same merging point are similar. If there exists a set of dominant merging points then the corresponding DRTTs would be dominant. It is well known that the AS topology has a power law degree distribution [6]. We expect that the number of prefixes sharing a merging point also follows similar distribution. To confirm this, we plotted in Fig.18 the fraction of prefixes that have the same merging point. The merging points are sorted according to the number of prefixes sharing them from the largest to the smallest. The graph is shown in log-log scale. It is obvious that the distribution follows power law. In particular, 30% of path pairs (thus prefixes) have a common merging point and therefore have similar DRTTs. Hence, it is not surprising that certain range of DRTTs are dominant.

So far we have seen the correlation between two paths at the same level such as two AS paths via the two providers. We now look at the correlation between IP level branch path pair and the AS level branch path pair. We refer to the merging point of two AS level paths as merging AS and similarly that of two IP level paths as merging IP. In general, the AS of the merging IP may not be the same as the merging AS because of the multiple peering points between ASes. Fig.19 shows whether the AS of the merging point of the IP level path is the same as the merging point of the AS level path. The x-axis represents the difference from the merging AS to the AS of the merging IP . The position difference 0 means that the merging AS is the same as the AS of the merging IP. If it is 1, the AS of the merging IP is the

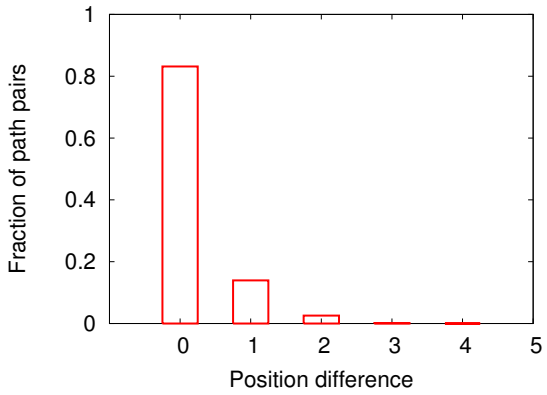

Figure 19: Merging AS vs AS of the Merging IP

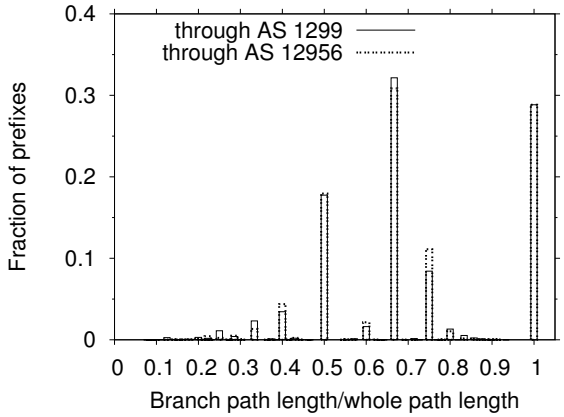

Figure 20: Ratio of the branch path length over the whole path length (Virtual Stub)

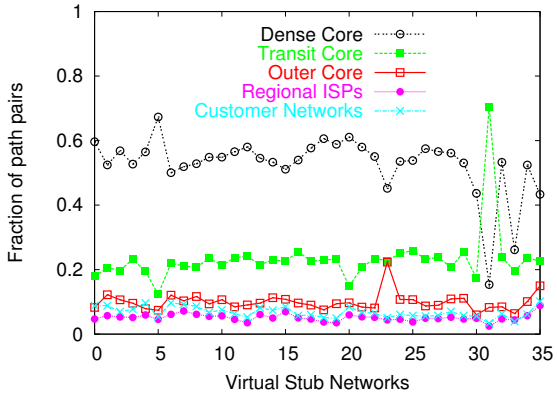

Figure 21: Fraction of AS path pairs that merge at each AS hierarchy position

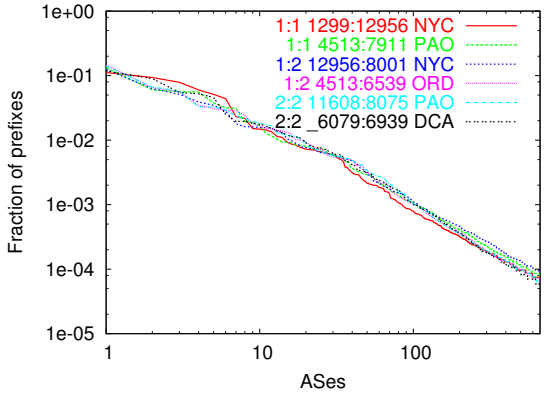

Figure 22: Merging Point Sizes (Virtual Stub)

next hop of the merging AS. As can be seen in Fig.19, for more than 83% of the destinations, the AS of the merging IP is the same as the merging AS. This implies that the conclusions drawn above would not be different whether the analysis is done on AS level or router level paths.

4.2 Analysis on the Routeviews Data

To verify that this path correlation phenomenon is very prevalent, we use Routeviews data as the model of other stub networks. Routeviews data contains a number of BGP routing tables [3]. We use the table of 'sh ip bgp' format RIBs collected on Oct 6th 2003. There are 10 cities that have more than one AS peering to Routeviews. We select 2 ASes from a city and consider them as the upstream providers of an imaginary stub network called *virtual stub network*. We call the 2 ASes *virtual providers*. This technique of an imaginary stub network was also used in [4] by treating two physically disjoint servers as if they were multihomed. Among the 10 cities, NYC and PAO have 5 and 7 ASes respectively and other cities have only 2 ASes. After we discard the ASes that have only small number of path information, we have 36 combinations of virtual stub networks. We consider the path information from the two virtual providers as the path information that the virtual stub network recieves from them.

Fig.20 shows the distribution of the ratio of the branch path length over the whole path length in a virtual stub network with the virtual providers AS 1299 and AS 12956 in New York. We choose the two virtual providers as a typical example of the virtual stub networks. Similar to the result of our stub network, the ratios with large fractions are 0.5, 0.667, 0.75 and 1. This suggests that the two upstream paths merge in the middle of the paths.

Where the two paths merge is also an interesting question. Fig.21 shows the fraction of path pairs merging at each hierarchy position, where the x-axis is numbered by the virtual stub networks. The hierarchy positions of some merging ASes are not identified, so the fraction may not be summed to 1. As can be seen in Fig.21, for most cases, the fractions of the path pairs that merge at the dense core are about 50 % to 70 %. This fraction of ASes that merge at the dense core is also similar to the result of our stub network. In conclusion, most of the upstream path pairs merge at the dense core.

Finally, we analyze the distribution of the fraction of the prefixes that each merging point deals with for the six representative virtual stub networks. As can be seen in Fig.22, similar to Fig.18, the distribution looks like a power law. The keys show the AS hierarchy positions, the AS numbers and the location of the two virtual providers. Other virtual stub networks show the similar results. One minor difference from our stub network is that in these cases, the largest frac-

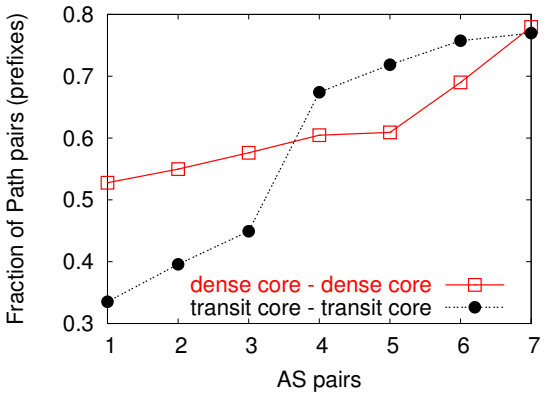

Figure 23: Fraction of equal length AS path pairs

tion is just a little more than 10 %. Nonetheless, the power law nature of the size of *branch path pair* remains same.

Another finding that has an impact on the best path selection procedure of the multihomed stub network is that the lengths of the two AS paths via the two providers are in many cases equal. This phenomenon is obvious when the two providers are in the same hierarchy position. Fig.23 shows the fraction of AS path pairs that have the same lengths. Each point represents one virtual provider pair. The virtual provider pairs are grouped by their hierarchy positions. The key shows the hierarchy positions of the two virtual providers. In many virtual stub networks, more than 50% of the AS path pairs have the same lengths. Especially when both the two virtual providers are in the dense core, the fractions are from 50 % to almost 80 %. This fact highly suggests that there should be an external mechanism to set the local preference properly, otherwise the next hop is always the provider with lower ID.

We conclude this section by summarizing that the virtual stub networks show very similar characteristics to our stub network in many aspects such as the location of the merging points and the dominance of a few merging points. Hence, we beleive that conclusions drawn based on the measurements from our stub network hold also for other stub networks.

5. IMPLICATIONS IN AS PATH SELECTION

Our findings have several important implications in intelligent provider selection using BGP. First, as analyzed in the previous section, for a large proportion of destination networks, there is considerable performance difference in using one provider over the other. This suggests that it is worthwhile to carefully select the best provider to reach certain destination networks for network latency optimization. Moreover, our results show that such provider selection can be done at AS-level for a large fraction of ASes (especially small and/or stub networks, who form the majority of the Internet). This makes the task of configuring appropriate routing policies for provider selection easier and less tedious. More importantly, our finding regarding the impact of AS hierarchy suggests that instead of performing RTT measurement to every single network prefix, we can exploit the correlation in RTTs to network prefixes with shared AS path segments to reduce the amount of measurement and monitoring traffic, thereby making intelligent provider selection for delay performance optimization more scalable.

Branch path pairs(Y,Z)		$I'(S,Y)$	$I'(S,Z)$	$pref'(S)$
Path via Y	Path via Z			
3356	209 3356	100	2	0.02
3356 701	209 701	143	101	0.7
3356 1239	209 701 1239	179	1	0.005
3356 7018	209 7018	334	0	0.0
3356 3549	209 701 3549	33	45	0.73
3356 209	209	0	2	0.0
3356 3561	209 3561	42	14	0.33
3356 2914	209 701 2914	68	2	0.029
3356 1668	209 701 1668	61	0	0.0

Table 3: Measurement to the merging points

Here we outline a simple algorithm for provider selection for delay optimization that does not require intrusive active probing to every destination network. To reduce the amount of probing traffic, it exploits the AS path correlation to adaptive sampling measurement. The algorithm is informally described as follows:

(1) Identify the set $\mathcal{C}$ of the branch path pairs in the AS level.

(2) Extract a path pair, P, from the set $\mathcal{C}$.

(3) Run the sampling measurement to the prefixes of the last AS of P.

(4) If the last AS of P shows a *preference meter* of smaller than a threshold, stop the measurement and assign high local preference to the preferred provider (more precisely, to the paths that contain the preferred provider) for those destinations containing P. Otherwise, expand P by adding one more hop to P and put the new path pairs into the set $\mathcal{C}$.

(5) If the set $\mathcal{C}$ is empty, stop. Otherwise go to the step (2).

In Table 3, an example of applying this algorithm to our stub network is shown. Table 3 shows the 9 branch path pairs and the initial sampling measurement. First and second column together form a branch path pair. Since the algorithm relies on sampling, we denote the estimated indicator function as "$I'(S,Y)$" and estimated preference function as "$pref'(S)$", respectively. The first iteration of the sampling measurement results are shown in columns $I'(S,Y)$, $I'(S,Z)$, and $pref'(S)$. Assuming the threshold is 0.05, we assign a high `Local Preference` value to the preferred provider for those prefixes that have the branch path pairs with $pref'(S)$ of less than 0.05, e.g., the branch pair "X Y 3356 & X Y 209 3356". For the next iteration, we expand the branch pairs with estimated preference higher than 0.05. After expanding each branch path pair, we repeat these steps. This algorithm can be used for automated `Local Preference` assignment to enable dynamic and intelligent provider selection.

6. CONCLUSION

In this paper we studied the potential benefits in choosing different providers for network latency optimization in a

multihomed stub network. Towards this end, we carried out extensive `ping` and `traceroute` measurements to a large set of destinations in a real multi-homed stub network. Based on the measurement data collected, we investigated the various factors that could potentially contribute to the performance differences in choosing different providers to reach various destination networks, in particular through correlation analysis among RTT measurements, BGP and traceroute data. Our study confirmed that there are considerable performance difference in terms of RTTs in choosing different providers to route traffic from a stub network to other destination networks. More importantly, we found that the Internet AS hierarchy has a strong impact on the RTT performance via different providers seen by a multi-homed stub network. This impact comes from the fact that the AS hierarchy often causes the AS paths via the different paths to merge at the core the Internet, resulting in shared common segments to many network prefixes and ASes. Consequently there is strong correlation among RTTs to many destination networks. This is manifested in our observation that for a large fraction of the network prefixes, the RTT differences between the two providers fall into a dominant range. We corroborated our findings on the AS path correlation due to the AS hierarchy through virtual stub network analysis using the Routeview BGP data. Our study sheds new light on the performance benefits of multi-homing and how such potential benefits can be exploited in an effective and scalable manner.

In particular, based on our findings, we outlined a simple procedure for choosing next-hop providers to optimize delay performance. It relies on selective performance sampling to reduce the amount of active measurement, and thus is potentially more scalable. We are currently evaluating and further refining the proposed provider selection algorithm. We also plan to extend our multi-homing study to multiple sites and with more providers, and to investigate other applications of our findings, e.g., in Internet distance mapping.

7. REFERENCES

[1] Proficient networks. http://www.proficient.net.

[2] Route science company. http://www.routescience.org.

[3] University of oregon route views project. http://www.routeviews.org.

[4] A. Akella, B. Maggs, S. Seshan, anees Shaikh, and R. Sitaraman. A measurement-based analysis of multihoming. In *Proceedings of ACM SIGCOMM 2003*, karlsruhe, Germany, Aug. 2003.

[5] A. Bremler-Barr, E. Cohen, H. Kaplan, and Y. Mansour. Predicting and bypassing end-to-end internet service degradations. In *Proceedings of Internet Measurement Workshop 2002*, Marseille, France, Nov. 2002.

[6] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of ACM SIGCOMM 1999*, Cambridge, MA, Sept. 1999.

[7] S. Halabi and D. McPherson. *Internet Routing Architectures*. CISCO PRESS, second edition, 2001.

[8] G. Huston. Interconnection, peering and settlements–part i. In *Internet Protocol Journal*, Jun 1999.

[9] Z. M. Mao, J. Rexford, J. Wang, and R. H. Katz. Towards an accurate as-level traceroute tool. In *Proceedings of ACM SIGCOMM 2003*, karlsruhe, Germany, Aug. 2003.

[10] N. Spring, R. Mahajan, and T. Anderson. Quantifying the causes of path inflation. In *Proceedings of ACM SIGCOMM 2003*, karlsruhe, Germany, Aug. 2003.

[11] L. Subramanian, S. Agarwal, J. Rexford, and R. H. Katz. Characterizing the internet hierarchy from multiple vantage points. In *Proc. IEEE INFOCOM*, New York, NY, June 2002.