A Measurement Study of Internet Bottlenecks

Ningning Hu*, Li (Erran) Li†, Zhuoqing Morley Mao‡, Peter Steenkiste* and Jia Wang§

* Carnegie Mellon University, Email: {hnn, prs}@cs.cmu.edu
† Bell Laboratories, Email: erranlli@dnrc.bell-labs.com
‡ University of Michigan, Email: zmao@eecs.umich.edu
§ AT&T Labs – Research, Email: jiawang@research.att.com

***Abstract*— Recent advances in Internet measurement tools have made it possible to locate bottleneck links that constrain the available bandwidth of Internet paths. In this paper, we provide a detailed study of Internet path bottlenecks. We focus on the following four aspects: the persistence of bottleneck location, the sharing of bottlenecks among destination clusters, the packet loss and queueing delay of bottleneck links, and the relationship with router and link properties, including router CPU load, router memory load, link traffic load, and link capacity. We find that 20% – 30% of the source-destination pairs in our measurement have a persistent bottleneck; fewer than 10% of the destinations in a prefix cluster share a bottleneck more than half of the time; 60% of the bottlenecks on lossy paths can be correlated with a loss point no more than 2 hops away; and bottlenecks can be clearly correlated with link load, while presenting no strong relationship with link capacity, router CPU and memory load.**

I. Introduction

Recent work has made it possible to identify the bottleneck link on Internet paths. An example is Pathneck [8]—a light-weight active probing tool that allows end users to identify the bottleneck location on a network path. Bottleneck location information is very useful for both Internet Service Providers (ISPs) and end users. ISPs can use it to locate network problems or to guide traffic engineering. End users can use it for server selection, multi-homing, and overlay routing, thus improving end-to-end performance.

However, before we can make intelligent use of bottleneck information, we need to gain a solid understanding of the properties of Internet bottlenecks. This includes the characterization of bottleneck link properties such as persistence, locality, path loss and queueing delay. A good understanding of these aspects will not only guide the measurement frequency for bottleneck monitoring tools, but it will also help network operators determine what kind of traffic engineering algorithms should be used to avoid bottlenecks. Furthermore, the understanding of bottleneck properties may provide insights in the causes of bottlenecks and their impact on network and end user performance.

In this paper, we answer the following questions. (i) What is the bottleneck location *persistence* over time? (ii) Do paths from a source to the destinations in the same network cluster share the same bottleneck? (iii) What is the relationship between bottleneck location and end-to-end path properties (e.g., packet loss rate and queueing delay)? (iv) What is the relationship between bottleneck location and router and link properties (e.g., routing change, link capacity/load, and router CPU and memory utilization)?

We use a Internet measurement study to address these questions. The bottleneck location information is obtained using Pathneck. The measurement sources and destinations are carefully selected to cover over 75,000 different Internet source-destination pairs. Some of these source-destination pairs are repeatedly measured for 38 days to study bottleneck persistence. To correlate bottlenecks with router and link properties, we obtain router and link statistics from a tier-1 ISP. Our main findings include the following. (i) On 20%–30% of the source-destination pairs in our measurements, the bottlenecks never change. (ii) For the end hosts within the same network prefix cluster, fewer than 10% of them share a bottleneck more than half the time. (iii) When correlating packet loss with bottleneck location, 60% of the bottlenecks on lossy paths can be correlated with a loss point no more than 2 hops away. (iv) Finally, a case study on a tier-1 ISP shows that the bottleneck location is clearly correlated with link load, while demonstrating no strong relationship with link capacity, router CPU and memory load.

The remainder of this paper is organized as follows. In the next section we briefly review the Pathneck tool and describe our data collection methodology. In Sections III, IV, and V, we look at bottleneck persistence, bottleneck sharing within network prefix clusters, and the relationship with loss rate and link queueing delay. Section VI provides a case study on a tier-1 ISP to reveal the relationship between bottleneck location and router and link properties. We discuss related work in Section VII and conclude in Section VIII.

II. Measurement Methodology

For each type of analysis, we use a variety of tools and methods to collect and analyze network measurement data. However, the Pathneck tool and the measurement sources and destinations selection method are used in all the studies we present. We discuss them in this section. For the convenience of reference, Table I lists the definition of the terms used in this paper.

A. Background on Pathneck

Pathneck is an active probing tool that allows end users to efficiently and accurately locate the bottleneck link on an Internet path. Pathneck is based on a novel probing technique called Recursive Packet Train (RPT) (Figure 1), which combines load and measurement packets. The load packets are UDP packets that are used to interact with background

TABLE I

TERMINOLOGY

Term	*Definition*
probing	the measurement using one RPT
probing set	n probings to the same destination; generally $n = 10$
persistent probing set	a probing set where all n probings follow the same route
choke point	a hop that limits the available bandwidth
bottleneck point	the last choke point on a path
location level	the routers in the same physical location are considered the same
AS level	the routers in the same AS are considered the same
dominant route	the most frequently used route by a path
route view	group the results based on route
end-to-end view	group the results based on source-destination pair

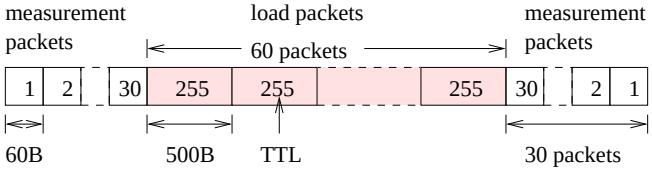


Fig. 1. Recursive Packet Train (RPT); the numbers in the boxes are TTL values

traffic and to obtain available bandwidth information. They are organized as a packet train, similar to the train used by end-to-end available bandwidth probing tools such as IGI/PTR [9] and Pathload [10]. The measurement packets, which precede and succeed the load packets as shown in Figure 1, are 60-byte UDP packets with the TTL fields set in such a way that at each hop along the path, the measurement packet at both the head and tail of the train will expire. This will trigger the transmission of two ICMP error packets to the source. The inter-arrival time (called the "gap value") of the ICMP packets at the source can be used as an estimate of the packet train length at the router that generated the two ICMP packets. The resulting sequence of packet train lengths at each hop can be used to identify the hop that limits the available bandwidth on the path. Hops where the packet train length increases have an available bandwidth that is lower than the packet transmission rate at that hop—we will call these hops *choke points*. The downstream link of a choke point is called a *choke link*. The last choke link is the bottleneck link.

In practice, queueing effects on both the forward and reverse paths and ICMP packet generation times can introduce noise in the train length measurements. To deal with this, Pathneck sends n consecutive RPTs (e.g., n = 10), called a *probing set*, and "averages" across these n probes. Only if a link repeatedly (e.g., more than half the probes) creates a significant increase in the train length (e.g., more than 10%) is it considered to be a valid choke point. This requirement is the main reason that Pathneck sometimes can not identify a bottleneck. The last choke point on the path is typically the link with the lowest available bandwidth, i.e., the bottleneck. The details of the algorithms can be found in [8].

Pathneck needs around 50 seconds to finish 10 probings. In our experiments for the persistence analysis, to guarantee each probing is conducted in a fixed amount of time, we allocate 90 seconds for each destination; this limits the number of destinations that each source can measure within a certain time interval. Besides bottleneck and choke point location, Pathneck also reports the IP address of each hop along the path, similar to the traceroute output. This information is used in this paper for route persistence analysis.

Pathneck is quite effective. An extensive Internet study [8] shows that it can detect bottlenecks for almost 70% of the paths. Pathneck also has relatively low overhead and does not require access to the destination. However, Pathneck does have some limitations. For example, it typically can not probe past firewalls since they often drop the load packets. Pathneck also cannot observe the last link of the path. For these reasons, the results presented in this paper are only for the partial paths for which we can obtain measurement data.

TABLE II

PROBING SOURCES FROM RON AND PLANETLAB (PL).

ID	Probing source	AS number	Location	Upstream provider(s)	Testbed	CL	IN
1	jfk1	3549	NY	1239, 7018	RON	√	√
2	lulea	2831	Sweden	1653	RON	√	√
3	ucsd	7377	CA	2152	RON	√	√
4	aros	6521	UT	701	RON	√	√
5	ana1	3549	CA	1239, 7018	RON	√	√
6	cornell	26	NY	6395	RON	√	
7	vineyard	10781	MA	209, 6347	RON	√	√
8	utah	17055	UT	210	RON	√	√
9	nyu	12	NY	6517, 7018	RON	√	√
10	ccicom	13649	UT	3356, 19092	RON	√	
11	nortel	11085	Canada	14177	RON	√	√
12	bkly-cs	25	CA	2150, 3356, 11423, 16631	PL		√
13	gr	3323	Greece	5408	RON		√
14	intel	7018	CA	1239	RON		√
15	mit-pl	3	MA	1	PL		√
16	princeton	88	NJ	7018	PL		√
17	purdue	17	IN	19782	PL		√
18	uga	3479	GA	16631	PL		√
19	umass	1249	MA	2914	PL		√
20	unm	3388	NM	1239	PL		√
21	uw-cs	73	WA	101	PL		√

"CL" denotes the measurements for clustering analysis;
"IN" denotes the measurements for router and link properties correlation.

B. Measurement Sources and Destinations

In our experiments, we run Pathneck from a host at Carnegie Mellon University and from a number of nodes selected from the RON and PlanetLab testbeds (listed in Table II). These nodes reside in 20 distinct ASes and are connected to 21 distinct upstream providers in north America and parts of Europe.

The measurement destination IP addresses are selected from BGP routing tables, as described in [18] and [8]. For the sources where we have local BGP tables, we directly use

them. Otherwise, we use the BGP tables from their upstream providers[1], which can be obtained from public BGP data sources such as Route Views [2]. The upstream provider information can be obtained by performing traceroute from the sources to a few randomly chosen locations such as `www.google.com`. Given a routing table, we first pick a ".1" or ".129" IP address for each prefix. The prefixes that are completely covered by their subnets are not selected. We then reduce the set of destination IP addresses by eliminating the ones whose AS paths starting from the probing source are completely covered by other AS paths. The motivation behind this is to achieve diverse AS-level coverage, while keeping the number of destinations manageable. The exact number of destinations selected depends on the goal of the analysis, as will be discussed as part of the methodology of each experiment. Note that the destination IP addresses obtained using this procedure do *not* necessarily correspond to real end hosts.

We did our best to diversify measurement sources and destinations so that our results can be as representative as possible. Even so, over half of our measurement sources directly connect to Internet-2, and the number of destinations is very small compared with the size of the Internet. For this reason, the conclusions drawn in this paper should not be viewed as representative of the whole Internet.

III. Persistence of Bottlenecks

In this section, we study the persistence of Internet bottlenecks. We first discuss our experimental methodology, and then look at route persistence. Finally, we discuss bottleneck persistence at various levels of spatial and temporal granularity.

A. Methodology

We study bottleneck persistence from both spatial and temporal perspectives. For the spatial analysis, we conducted *1-day periodic probing*. That is, we selected a set of 960 destinations and probed each of them once per day from a CMU host for 38 days. That provides us 38 sets of probing results for each destination. Here the number of destinations–960–is determined by the length of the probing period (1 day) and the measurement time of Pathneck (90 seconds per destination). This set of data is used throughout this section.

For the temporal analysis, we conducted two more experiments: (1) *4-hour periodic probing*, where we select a set of 160 destinations from those used in the 1-day periodic probing and probe each of them from a CMU host every four hours for 148 hours, obtaining 37 sets of probing results for each destination; and (2) *1-hour periodic probing*, where we select a set of 40 destinations from those used in the 4-hour periodic probing and probe each of them from a CMU host every hour for 30 hours, thus obtaining 30 sets of probing results for each destination. These two data sets are only used in Section III-D.

[1]In the case of multihomed source networks, we may not be able to obtain the complete set of upstream providers.

TABLE III
Determining co-located routers

Heuristic	*# IP pairs*
Same DNS name	42
Alias	53
CMU or PSC	16
Same location in DNS name	572
Digits in DNS name filtered	190
Real change	1722

B. Route Persistence

Bottleneck location can change when the underlined route changes. In the 1-day periodic probing data set, we observe quite a few IP level route changes: among the 6,868 unique IP addresses observed in this data set, 2,361 of them are associated with hops whose IP address changes, i.e., the route appears to change. This shows that we must consider route persistence in the bottleneck persistence analysis. Intuitively, Internet routes have different persistence properties at different granularity, so in the following, we investigate route persistence at both the location and AS level. At the *location level*, we consider hops with IP addresses that belong to the same router or co-located routers as the same hop. We will explain what we mean by the "same router" or "co-located router" below. Location-level analysis can help us reduce the impact of "false" route changes. At the *AS level*, we consider all hops in the same AS as the same AS-level hop; this is done by mapping the IP address of each hop to its AS number using the mapping provided by [18].

1) Location-Level Route: At the location level, the IP addresses associated with the same router are identified using two heuristics. First, we check the DNS names. That is, we resolve each IP address into its DNS name and compare the DNS names. If two IP addresses *(a)* have the same hop position *(b)* for the same source-destination pair and *(c)* are resolved to the same DNS name, they are considered to be associated with the same router. We found that 5,410 out of the 6,868 IP addresses could be resolved to DNS names, and 42 pairs of IP addresses resolve to identical DNS names (refer to Table III). Second, we look for IP aliases. For the unresolved IP addresses, we use Ally [22] to detect router aliases. We found that 53 IP pairs are aliases.

The IP addresses associated with co-located routers are identified by applying the following heuristics sequentially.

1) *CMU or PSC*. Because all our measurements are conducted from a CMU host, they always pass through PSC (`www.psc.edu`) before entering other networks, so we consider all those routers within CMU or PSC as co-located.
2) *Same location in DNS name.* As pointed out in [25], the DNS names used by some ISPs (e.g., the `*.ip.att.net` for AT&T and the `*.sprintlink.net` for Sprint) include location information, which allows us to identify those routers that are at the same geographical position.
3) *Digits in DNS name filtered*. We remove the digits from

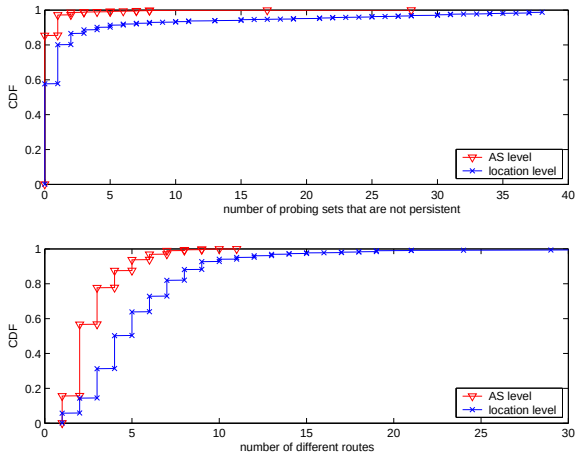


Fig. 2. Route persistence at the location level and AS level

DNS names. If the remaining portion of the DNS names become identical, we consider them to be co-located.

These three heuristics allow us to identify 16, 572, and 190 pairs of co-located routers, respectively. Note that heuristics (2) and (3) are not perfect: stale information in DNS can cause mistakes in heuristic (2), while heuristic (3) is completely based on our limited knowledge of how ISPs assign DNS names to their IP addresses. Although we think the impact from these errors is small, better tools are needed to identify co-located IP addresses.

At the location level, we consider a route change only when the corresponding hops do not belong to the same or a co-located routers. Table III shows that 1,722 pairs of IP addresses are associated with hops that experience route changes. Given this definition for location-level route change, we define a *persistent probing set* as a probing set where the route remains the same during the 10 probings.

2) Results: Figure 2 shows the route persistence results for the 1-day periodic probing, at both the location and AS level. The top graph plots the cumulative distribution of the number of probing sets that are not persistent. As expected, AS-level routes are more persistent than location-level routes. Some location-level routes change fairly frequently. For example, about 5% of the source-destination pairs have more than 15 (out of 38) probing sets that are not persistent at the location level. However overall, the vast majority of the routes are fairly persistent in the short term: at the location level, 57% of the source-destination pairs have perfect persistence (i.e., all probing sets are persistent), while 80% have at most one probing set that is not persistent. The corresponding figures for AS level are 85% and 97%, respectively.

The bottom graph in Figure 2 illustrates long-term route persistence by plotting the distribution of the number of different location-level and AS-level routes that a source-destination pair uses. We observe that only about 6% of the source-destination pairs use one location-level route, while about 6% of the source-destination pairs have more than 10 location-level routes (for 380 probings). The long-term route

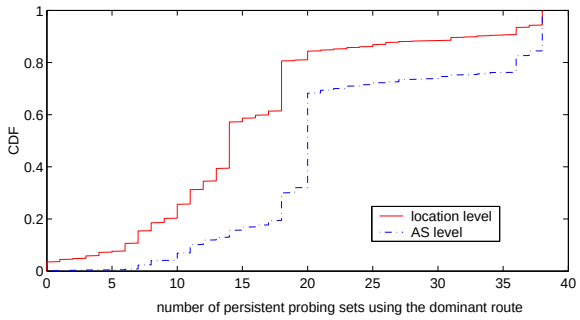


Fig. 3. Frequency of the dominant route

persistence at the location level is quite poor. However, at the AS level, not surprisingly, the routes are much more persistent: 94% of the source-destination pairs have fewer than 5 different AS-level routes.

We have seen that most of the source-destination pairs use more than one route. For our bottleneck persistence analysis, we need to know if there is a dominant route for a source-destination pair. Here, the *dominant route* is defined as the route that is used by the highest number of persistent probing sets in all 38 probing sets for the same source-destination pair. Figure 3 shows the distribution of the dominant route for each source-destination pair, i.e., the number of persistent probing sets that use the dominant route. We can see that, at the location level, only around 15% of the source-destination pairs have a route with a frequency of 20 or more (out of 38), i.e., the "dominant" routes are usually not very dominant. At the AS level, for about 30% of the source-destination pairs, the dominant route is used by less than 20 (out of 38) probing sets. This is consistent with the observation in [25] that a total of about 1/3 of Internet routes are short lived (i.e., exist for less than one day).

C. Spatial Bottleneck Persistence

We study spatial bottleneck persistence from two points of view: the route view and the end-to-end view. The route-view analysis provides the bottleneck persistence results excluding the effect of route changes, while end-to-end view can tell us the bottleneck persistence seen by a user, including the effect of route changes. The comparison between these two views will also illustrate the impact of route changes. In each view, the analysis is conducted at both the location and the AS level. A bottleneck is persistent at the location level if the bottleneck routers on different routes for the same source-destination pair are the same or co-located. A bottleneck is persistent at the AS level if the bottleneck routers on different routes for the same source-destination pair belong to the same AS.

1) Route View: In the route view, bottleneck persistence is computed as follows. We first classify all persistent probing sets to the same destination into different groups based on the route that each probing set follows. In each group, for every bottleneck router detected, we count the number of persistent probing sets in which it appears (cnt), and the number of persistent probing sets in which it appears as a bottleneck

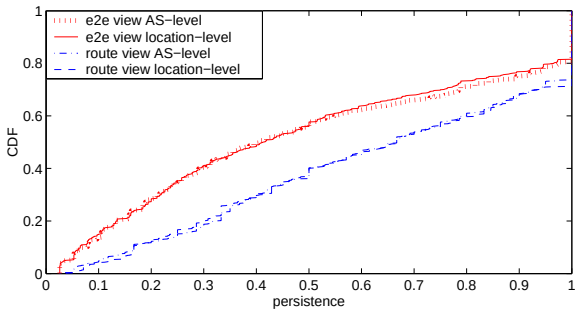


Fig. 4. Persistence of bottlenecks.

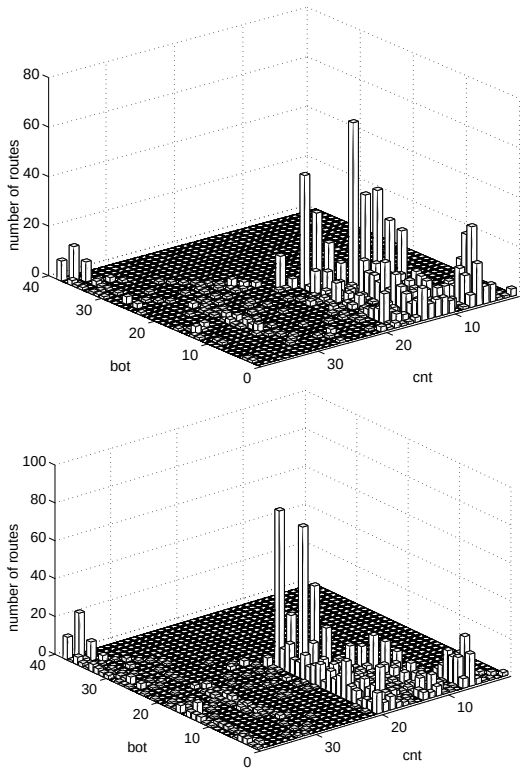


Fig. 5. Number of routes with certain (bot, cnt) value at the location level (top) and AS level (bottom).

(bot). Then the bottleneck persistence is defined as bot/cnt. To avoid the bias due to small cnt, we only consider those bottlenecks where $cnt \geq 10$. The number "10" is selected based on Figure 3, which shows that over 80% (95%) of the source-destination pairs have a dominant route at the location level (AS level) with a frequency higher than 10; also, picking a larger number will quickly reduce the number of source-destination pairs that can be used in our analysis. Therefore, 10 is a good trade-off between a reasonably large cnt and having a large percentage of source-destination pairs to be used in the analysis.

In Figure 4, the two bottom curves (labeled with "route view") plot the cumulative distribution of the bottleneck persistence. We can see that, at both the location level and AS level, around 50% of bottlenecks have persistence larger than 0.7, and over 25% of them have perfect persistence. This shows that most of the bottlenecks are reasonably persistent in the route view. Note that the location-level curve and the AS-level curve are almost identical. This seems to contradict the intuition that bottlenecks should be more persistent at the AS level. Note however that for a source-destination pair, cnt in the AS level can be larger than that for the location level, so we cannot directly compare the persistence at these two levels.

In Figure 5, we look at the route-view persistence in more detail by plotting the number of bottlenecks falling into each (bot, cnt) category. The results for the location level (top) and AS level (bottom) are fairly similar. We observe that most of the routes cluster in the triangular region within $0 < bot < 20 \ \& \ 0 < cnt < 20$. This is not surprising, since it reflects the fact that many routes for a source-destination pair appear in fewer than 20 of the daily probings. An important message is that there is a higher concentrations of bottlenecks close to the diagonal, suggesting that bottlenecks are fairly persistent.

2) End-To-End View: In this view, we consider bottleneck persistence in terms of source-destination pairs, regardless of the route taken. We compute bottleneck persistence of end-to-end view in a way similar with that of route view. The two top curves (labeled with "e2e view") in Figure 4 show the results for end-to-end bottleneck persistence. Again, the results for location level and AS level are very similar. However, the persistence in the end-to-end view is much lower than that in the route view – only 30% of bottlenecks have persistence larger than 0.7. This degradation from that in the route view illustrates the impact of route changes on bottleneck persistence.

Figure 6 plots the distribution details. Similar to the route-view results shown in Figure 5, the location-level and AS-level results are very similar to each other. However, there is an obvious difference — most source-destination pairs are clustered in the area $30 < cnt \leq 38$, which reflects the fact that for each source-destination pair we have 38 probing sets. Here cnt can be less than 38 because we only consider persistent probing sets. Comparing with the results in Figure 5, we can see that route changes can easily change the end user's perception of bottleneck persistence.

3) Relationship With Gap Values: For those bottlenecks with high persistence, we find that they tend to have large gap values in the Pathneck measurements. This is confirmed in Figure 7, where we plot the relationship between the bottleneck gap values and their persistence values in both the route view and end-to-end view. We split the bottlenecks that are included in Figure 4 into 4 groups based on their persistence value: 1, [0.75, 1), [0.5, 0.75), and [0, 0.5), and then plot the cumulative distribution for the average bottleneck gap values in each group. We observe a clear relationship between large gap values and high persistence in both the route view (top figure) and end-to-end view (bottom figure). The reason is, as discussed in [8], that a larger gap value corresponds to smaller available bandwidth, and the smaller the available bandwidth, the less likely it is that there will be

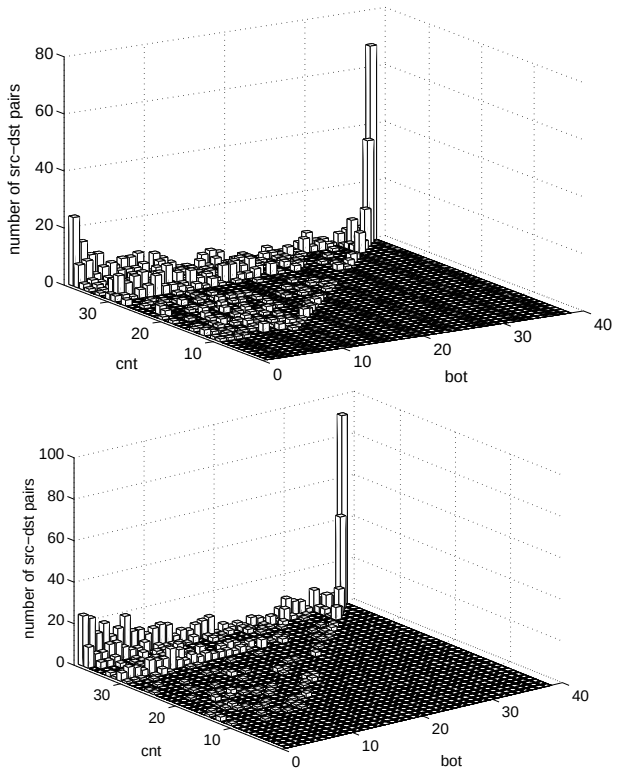


Fig. 6. Number of source-destination pairs with certain (bot, cnt) value at the location level (top) and AS level (bottom).

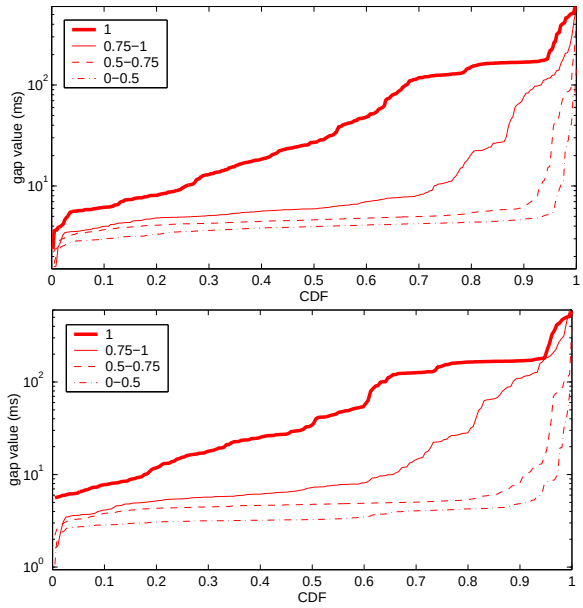


Fig. 7. Bottleneck gap value vs. bottleneck persistence for both the route view (top) and end-to-end view (bottom).

a hop with a similar level of available bandwidth on the path between a source-destination pair, so the bottleneck is more persistent.

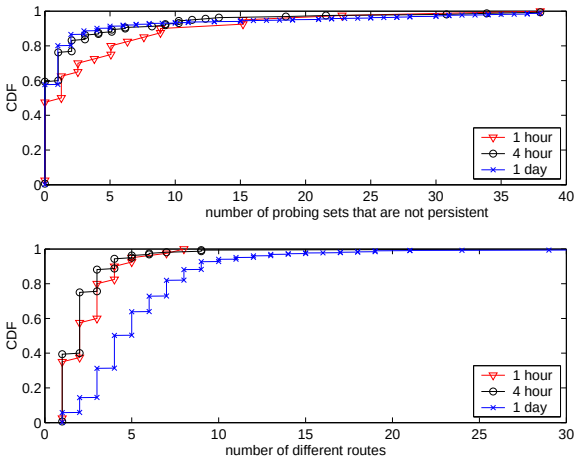


Fig. 8. Location-level route persistence.

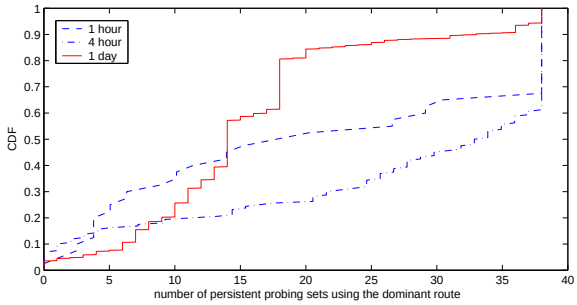


Fig. 9. Distribution of the dominant route at the location level.

D. Temporal Bottleneck Persistence

So far our analysis has focused on the 1-day periodic probing results, which provide only a coarse-grained view of bottleneck persistence. The 4-hour and 1-hour periodic probings described early in this section allow us to investigate short-term bottleneck persistence. Although these two sets of experiments only cover a small number of source-destination pairs, it is interesting to compare their results with those in the 1-day periodic probings.

Figure 8 compares location-level route persistence over 1-hour, 4-hour, and 1-day time periods. In the top graph, the x-axis for the 1-hour and 4-hour curves are scaled by 38/30 and 38/37 to get a fair comparison with the 1 day curve. For the 4-hour and 1-day periodic probings, the number of probing sets that are not persistent are very similar, while those for 1-hour periodic probing show a slightly higher percentage of probing sets that are not persistent. This seems to imply that there are a quite a few short-term route changes that can be caught by 1-hour periodic probings but not by 4-hour periodic probings. The bottom figure shows that the number of different routes for 1-day periodic probings is significantly larger than those for 4-hour and 1-hour periodic probings. We think this is mainly because the 1-day periodic probings cover a much longer period.

Figure 9 plots the distribution of the dominant route at the location level. Clearly, in the 4-hour and 1-hour periodic

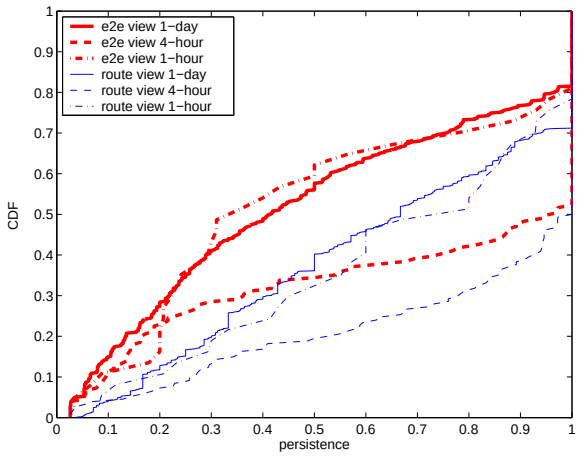


Fig. 10. Persistence of the bottlenecks with different measurement periods at the location level.

probings, the dominant routes cover more persistent probing sets than for the 1-day periodic probings — in the 4-hour and 1-hour periodic probings, 75% and 45% of the source-destination pairs have over 20 persistent probing sets that use the dominant routes, while only around 20% of the source-destination pairs in the 1-day periodic probings use the dominant routes. Note that the 4-hour periodic probing results have the largest dominant route coverage. A possible reason is that the 1-day periodic probings last much longer and allow us to observe more route changes, while the 1-hour periodic probings can catch more short-term route changes. The same explanation can also explain the difference in bottleneck persistence plotted in Figure 10, which compares the location-level bottleneck persistence for different probing periods. Again, we see that the 1-day and 1-hour curves are closer to each other in both the route view and the end-to-end view, while the 4-hour curves stand out distinctly, with higher persistence. This is because the 4-hour periodic probings have the best dominant route coverage, so route changes have the least impact.

E. Summary of Bottleneck Persistence Study

The analysis in this section shows that 20% – 30% of the bottlenecks have perfect persistence. As expected, bottlenecks at the AS level are more persistent than bottlenecks at the location level. Long-term Internet routes are not very persistent, which has a significant impact on the bottleneck persistence. That is, people will reach different conclusions about bottleneck persistence depending on whether or not route changes are taken into account. We also confirm that bottlenecks with small available bandwidth tend to be more persistent. Finally, we show that bottleneck persistence is also sensitive to the length of the time period over which it is defined, and the worst persistence results seem to occur for medium time periods.

IV. Choke Link Sharing In Destination Clusters

In this section, we investigate the degree of choke link sharing among paths from a probing source to destinations whose IP addresses are within the same network cluster. As defined by Krishnamurthy and Wang [13], a network cluster is a set of nodes that share the same prefix in the BGP routing table. Previous work by Balakrishnan *et al.* [4] has found that Internet hosts close to each other often have similar throughput. The goal of our study is to understand whether the network paths from a randomly selected probing source to destinations close to each other experience the same bandwidth choke links. Note that we are interested in the entire set of choke links, not just bottlenecks. Such information can be very valuable in reducing unnecessary probing, producing more accurate performance prediction, or in general in doing performance-based clustering of IP addresses.

There are several reasons why we cannot always expect destinations whose IP addresses are within the same prefix to have the same choke links from the perspective of a given vantage point. First of all, network paths from the probing source to those destinations may not necessarily follow the same AS-level paths due to reasons such as BGP misconfiguration and address aggregation [18]. Second, even assuming the AS-level paths are the same, the IP level paths can disagree resulting in different choke links. Finally, the choke link locations may not be persistent, resulting in different path characteristics to destinations within the same cluster. To understand the degree of choke link sharing for end hosts within a cluster network, we conduct the following study.

A. Methodology

We use 11 probing sources, all RON nodes, as shown in column "CL" in Table II, to collect the measurement data. To reduce the bias caused by not discovering the last mile bottleneck, we intentionally selected addresses from a large set of local DNS server IP addresses as target addresses. In addition, we ensure that all the selected IP addresses are responsive to ICMP ping requests, so that it is more likely that Pathneck can successfully probe the last several hops of the network path. To have conclusive results, we select 20 to 60 IP addresses belonging to each prefix cluster. As a result, for each probing source, 1087 IP addresses are selected; they belong to a diverse set of prefixes originating from ASes across the entire Internet hierarchy. The measurements from all probing sources were conducted roughly around the same time—the starting times are within a 60-second interval, and the ending times are within a 60-minute interval.

B. Choke Link Sharing Within A Prefix

Figure 11 shows the results across all 11 probing sources. The *Degree of Sharing* is calculated as the percentage of the paths (from a source to destinations in the same network cluster) in which the most popular choke link occurs. The figure shows the correlation at three levels of granularity: IP level, location level, and AS level. First, we observe that more than 80% of the prefixes have only at most 20% sharing for

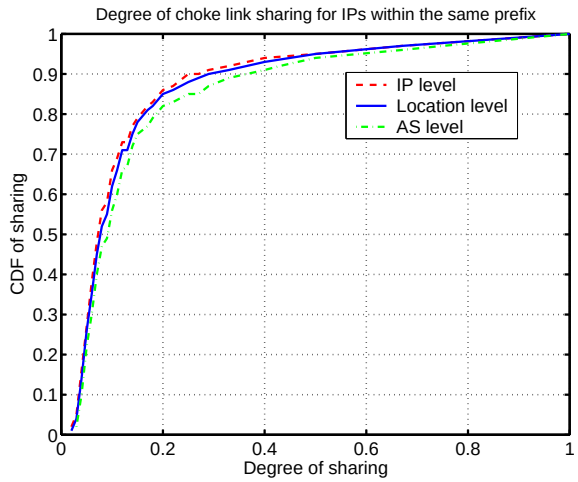


Fig. 11. Degree of choke link sharing at different levels for IPs within the same prefix

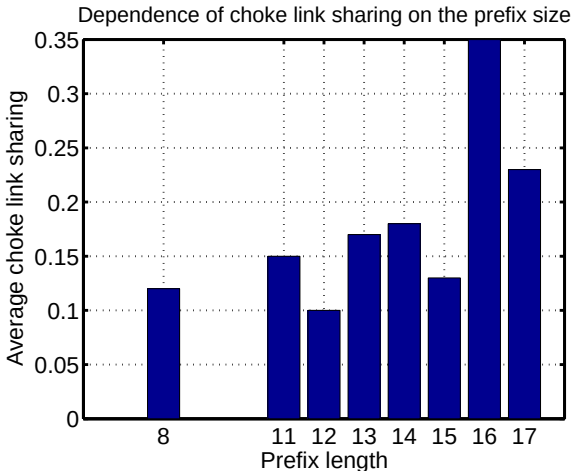


Fig. 12. Impact of prefix length on the degree of sharing

the IP addresses selected, and about half of the prefixes have at most 10% sharing. Second, there is a slight improvement from the IP level when either location or AS level correlation is used although the difference is negligible. However, we did find that in most cases, even though the network paths to the same prefix have different choke links, the difference in their position is only 1 or 2 hops. We also looked at the degree of sharing using the measurements from each individual source. We did not observe significant differences across different probing sources.

One explanation for the low degree of choke link sharing within the same prefix is the large size of the prefix. Address aggregation can merge groups of smaller prefixes into a large prefix. Such smaller prefixes within a large prefix may not follow the same AS level path, so they may not share choke links. To test this hypothesis, we study the impact of address prefix length on the degree of sharing in Figure 12. We observe that as the prefix length increases, the degree of sharing also tends to increase though not in a consistent way.

As part of the future work, we plan to probe more extensively to better understand why the degree of choke links sharing from a source to a destination cluster is very small, and to further validate our conjecture that aggregation plays a role in choke link sharing.

V. Relationship With Link Loss And Delay

In this section, we investigate whether there is a clear relationship between bottleneck and link loss and delay. Since network traffic congestion may cause queueing, packet loss and hence bottlenecks, we expect to see that bottleneck points are more likely to experience packet loss and queueing delay. On the other hand, capacity determined bottlenecks may not experience packet loss. Therefore, the relationship between bottleneck position and loss position may help us to distinguish load-determined and capacity-determined bottlenecks.

TABLE IV

Different types of paths in the 954 paths probed

	No loss	Loss	Total
No bottleneck	139	121	260
Bottleneck	312	382	694
Total	451	503	954

In this study, we use Tulip [17] to detect the packet loss position and estimate ling queueing delay. We probed 954 destinations from a CMU host. For each destination, we did one set of Pathneck probings, i.e., 10 RPT probing trains, followed by a Tulip loss probing *and* a Tulip queuing probing. Both types of Tulip probings are configured to conduct 500 measurements for each router along the path [1]. For each router along the path, Tulip provides both the round trip loss rate and forward path loss rate. Because Pathneck can only measure forward path bottlenecks, we only consider the forward path loss rate. Table IV classifies the paths based on whether or not we can detect loss and bottleneck points on a path.

A. Relationship with Loss

Let us first look at how the positions of the bottleneck and loss points relate to each other. In Figure 13, we plot the distances between loss and bottleneck points for the 382 paths where we observe both a bottleneck and loss points. In the top figure, the x-axis is the normalized position of a bottleneck point — the normalized position of a hop is defined to be the ratio between the hop index (the source node has index 1) and the length of the whole path. The y-axis is the relative distance from the closest loss point to that bottleneck point. If there is a loss point with equal distance on each side, we plot both, one with a positive distance, and the other with a negative distance. Positive distance means that the loss point has a larger hop index, i.e., it is downstream from the bottleneck point; negative distance means that the loss point is earlier in the path than the bottleneck point. The bottom figure presents the data from

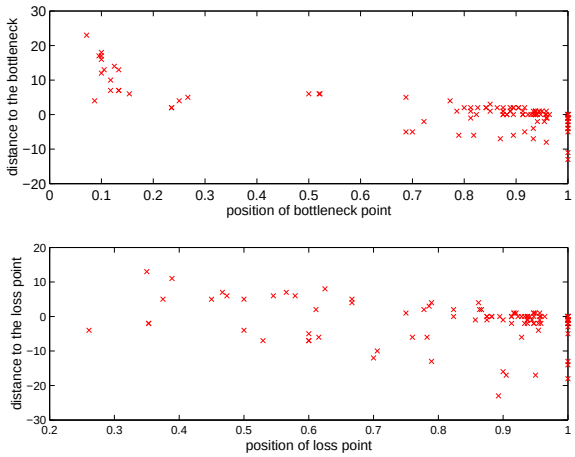


Fig. 13. Distances between loss and bottleneck points.

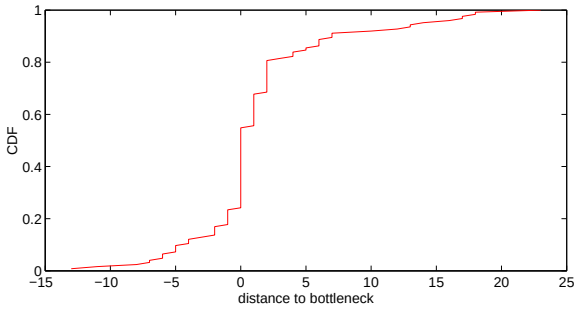


Fig. 14. Distance from closest loss point to each bottleneck points

the loss point of view, and the distance is computed from the closest bottleneck point. Figure 13 clearly shows that there are fewer bottleneck points in the middle of the path, while a fair number of loss points appear within the normalized hop range [0.3, 0.9]. On the other hand, there are fewer loss points in the beginning of the path.

Figure 14 shows the cumulative distribution of the distance from the closest loss point to each bottleneck points, using the same method as that used in the top graph of Figure 13. We observe that over 30% of bottleneck points also have packet loss, while around 60% of bottleneck points have a loss point no more than 2 hops away. This distance distribution skews to the positive side due to the bottleneck clustering at the beginning of the path, as shown in Figure 13.

B. Relationship with Delay

Besides packet loss, queueing delay is another metric that is frequently used as an indication of congestion. Tulip provides queueing delay measurements as the difference between the median RTT and the minimum RTT from the probing source to a router. Note that the queueing delay computed this way corresponds to the cumulative queueing delay from the probing source to a router, including delay in both the forward and return path. The 500 measurements for each router in our experiment can provide a reasonable estimate for this

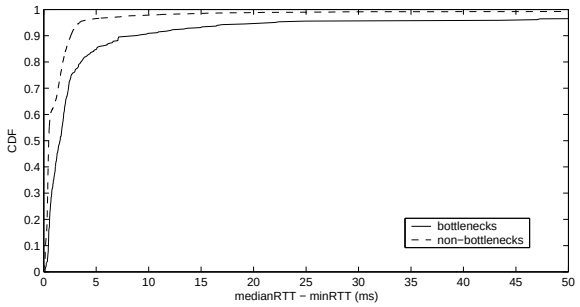


Fig. 15. Bottlenecks vs. queueing delay

queueing delay. Based on these measurements, we look at the relationship between the bottlenecks and the corresponding queueing delays.

Figure 15 shows the cumulative distribution of the queueing delays for bottleneck and non-bottleneck links. In our experiment, we observe queueing delays as large as $900ms$, but we only plot up to $50ms$ in the figure. As expected, we tend to observe longer queueing delays at bottleneck points than at non-bottleneck points: fewer than 5% of the non-bottleneck links have a queue delay larger than $5ms$, while around 15% of the bottleneck links have a queue delay larger than $5ms$. We also observe the same relationship between the loss points and their queueing delays, i.e., we tend to observe longer queueing delay at the loss points.

VI. Impact of Router and Link Properties

In this section, we use information obtained from the network of a tier-1 ISP X to study various factors behind observed choke links or bottlenecks in the forwarding path segments that traverse X. These factors include router CPU load, router memory load, link capacity and link load. Below, we first describe how we use end-to-end probing to cover links inside X, and how we identify inter/intra-AS links. We then present the relationship between choke links and the corresponding router and link performance properties.

A. Covering ISP X

Ideally, we would like to run Pathneck for paths connecting each pair of ingress and egress interfaces of X, identify choke links and bottlenecks on each path, and then investigate the causes for the choke links and bottlenecks. Unfortunately, we do not have direct access to the ingress and egress points. Instead, we use probing sources outside of ISP X, and carefully choose a large set of destinations for each probing source to cover as many distinct inter-AS links as possible that connect to X. As a result, the probing paths can also traverse a large number of distinct intra-AS links within X. Specifically, we choose 19 RON and PlanetLab nodes as the probing sources as listed in the column “IN” of Table II. Due to their different positions in the Internet, they cover different numbers of inter-AS and intra-AS links of X. As a result, the number of probing destinations selected for each probing source is different – it varies from around 800 to around 8,000.

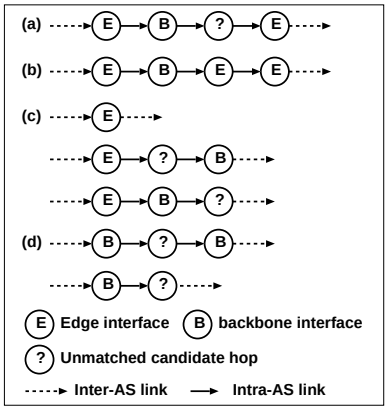


Fig. 16. Identify inter- and intra-AS links of X on a probing path.

In total, we collected a total number of 66,876 probing sets, each containing 10 consecutive probings.

Our method of selecting measurement paths maximizes the coverage of X, but it does not guarantee that the bottlenecks are in X. First, the choke links and bottlenecks may be outside of ISP X. Second, due to route changes, some pre-determined probing paths may not traverse X at all when we conducted the measurements, so they do not cover any links within X.

B. Identify Links Belonging to ISP X

For our analysis, we need to identify the path segment that is within ISP X. In general, identifying the segment of a path that traverses an arbitrary AS X is very hard [8]. Tools such as Traceroute and Pathneck return one IP address for each router (hop) along the path between a source and destination. Simply mapping this IP address to its AS number and identifying the hops with AS number X might not yield a correct result due to the naming convention adopted by some ISPs. For example, an inter-AS link could have two end IP addresses belonging to the same AS. Fortunately, we have access to the router configuration files of all the edge routers and backbone routers in ISP X. We parse these configuration files, extract the IP addresses of all the interfaces, and group the interface IP addresses into *edge interfaces* (interfaces that connect to a router in a neighboring network) and *backbone interfaces* (interfaces that connect to a router in ISP X). Given this information, we first map the IP address of each hop along a path to its AS number and identify all the hops with AS number X and their adjacent hops as the candidate hops. We then match the IP address of each candidate hop with the edge interface addresses and the backbone interface addresses. Based on this classification, we use the following heuristics to identify the inter-AS and intra-AS links.

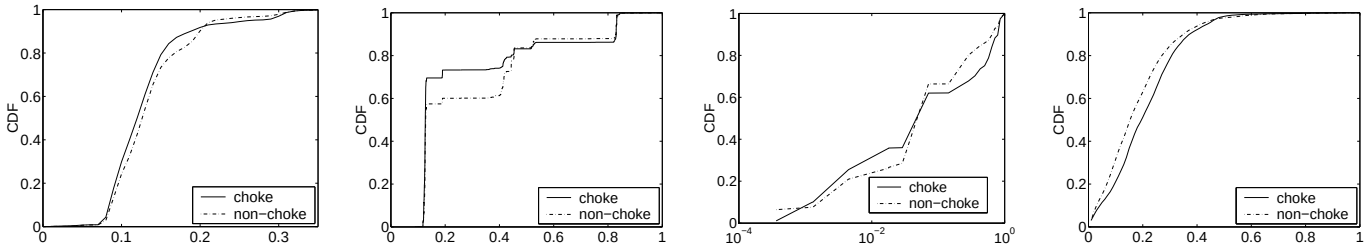
1) If we identify two hops on a path as edge interfaces (Figure 16(a)), we consider the links between these two hops as intra-AS links of X. The two end links, i.e., the link preceding the first edge interface and the link following the second edge interface are considered as inter-AS links.
2) If we identify more than two hops on a path as edge interfaces (Figure 16(b)), we consider the first and the last edge interfaces as the "real" edge interfaces and apply Heuristics 1).
3) If we can only identify one edge interface on a path (Figure 16(c)), we must consider several cases. If this edge interface is the only candidate hop, then we consider its two adjacent links as inter-AS links of X and there is no intra-AS link. If there is more than one candidate hop and at least one backbone interface is also identified, we consider the following two cases. If the last candidate hop is identified as a backbone interface, we consider this backbone interface as an edge interface. If the last candidate hop does match with any address we have and it is adjacent to a backbone interface, we consider this unmatched candidate hop as an edge interface and apply Heuristics 1).
4) If no edge interfaces are identified (Figure 16(d)), we consider the following two cases. If the first and the last candidate hops are identified as backbone interfaces, we consider them as edge interfaces. If the first (or the last) candidate hop is unmatched and it is adjacent to a backbone interface, we consider it as edge interface and apply Heuristics 1).

After applying the above heuristics to our probing data set, we get 429,908 "valid" probings. Among them, we identified 7,641 distinct links related to ISP X, among which 3,419 links are intra-AS links and 4,222 links are intra-AS links.

C. Location of Choke Links

With an accurate identification of inter-AS and intra-AS links, we now validate the common belief that bottlenecks are more likely to be on the inter-AS links, including peering and access links. Due to the limitations of our data collection method mentioned earlier, we are unable to conduct a meaningful study on the bottleneck link, because the vast majority of the detected bottlenecks are outside of ISP X. This is not surprising because ISP X is a well-engineered tier-1 service provider. In the following analysis, we study the location of the choke links detected in ISP X. We use the *detection rate* to measure how likely a link appears as a choke link on a path. The detection rate is defined as the number of times that a link is detected as a choke link divided by the number of times the link appears in the probing paths.

Figure 17 shows the cumulative distribution of detection rate of inter-AS and intra-AS links. We observe that, in ISP X, inter-AS links are much more likely to be choke links than intra-AS links — only around 5% of the intra-AS links have detection rates larger than 0.3, while around 30% of inter-AS links have a detection rate over 0.3. This is consistent with the common belief that the bottlenecks are likely to be on the peering and access links. However, note that a choke link (or even bottleneck link) does not directly correspond to congestion in a network. In fact, based on the packet loss and link load information, we did not observe any congestion in ISP X during the period we conducted our experiments. In



(a) Router CPU load (b) Router memory utilized (c) Link capacity (d) Link load

Fig. 18. CDFs of router CPU load, memory load, link capacity (normalized), link load for choke links and non-choke links in ISP X.

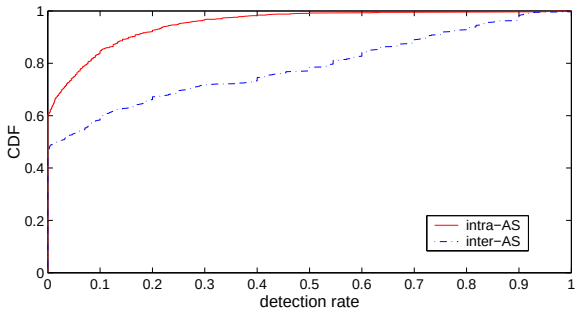


Fig. 17. Detection rate of the inter-AS and intra-AS links.

addition, the conclusion on choke links detected in ISP X may not apply to choke links (and bottleneck links) in other networks such as smaller ISPs or enterprise networks.

D. Causes of Choke Links

We further investigate the various factors that may cause a link to be a choke link. We consider the following two factors: router utilization and link utilization. We use four metrics in our analysis: *router CPU load*, *router memory load*, *link capacity*, and *link load*. These data are obtained from the 5-minute SNMP statistics collected from ISP X's internal routers. Note that other factors may affect choke links such as packet loss and routing changes. In our analysis, we do not consider packet loss because it rarely happens in ISP X; we will investigate the impact of route changes on choke links as part of future work.

Our conjecture is that the link capacity and traffic are major factors behind the choke links, while router performance has less impact. The intuition is that the packet forwarding processing is mostly done on the line cards [7]. We validate our conjecture below.

We do not observe a strong correlation between the router CPU/memory utilization and the probability of the router being a choke point, mainly due to the light load on all routers. Figure 18(a) shows the cumulative distributions of router CPU load for choke routers and non-choke routers in ISP X. The CPU load on all the routers traversed by our Pathneck probes is lower than 35%. Similarly, router memory utilization is also low as illustrated in Figure 18(b). These results confirm our conjecture that router CPU and memory load do not affect the likelihood of being a choke point.

Second, there is no clear relationship between link capacity and the probability of being a choke point. Figure 18(c) shows the cumulative distributions of normalized link capacity for choke links and non-choke links. Intuitively, one may think low capacity links are likely to be choke links. However, we observe that high capacity links have similar probability of being a choke link as the low capacity links. This might be due to the fact that the network is engineered such that traffic load is well-balanced according to the link capacities.

Finally, we do observe a correlation with the link load. Figure 18(d) shows the cumulative distributions of normalized link load for choke links and non-choke links. Choke links have a slightly higher link load than non-choke links. Though the correlation between link load and its probability of being a choke link does not directly tell us what the causes of choke links are, it provides hints that traffic load might be one of the major factors that cause Internet bottlenecks.

VII. Related Work

Our work studies the persistence of bottlenecks of Internet paths, the extent of bottleneck sharing among IP addresses in destination clusters, the correlation of different path properties, and the relationships between choke links and router and link properties. We review related work on each of these four topics.

Persistence of Internet path properties. To our knowledge, the persistence of Internet bottleneck locations has not been well explored. However, the persistence of other Internet path properties have been investigated in the literature. These include control path (BGP route) and forwarding path persistence, path loss, packet ordering, path delay, and throughput. Labovitz *et al.* [14], [15], [16] showed that a large fraction of destination prefixes have persistent routes from many observation points despite the large volume of BGP updates. Rexford *et al.* discovered that the small number of popular destinations responsible for the bulk of Internet traffic have very persistent BGP routes. Zhang *et al.* [25], [24] investigated the stationarity of forwarding path, loss and throughput. They show that routes appear to be very persistent although some routes exhibited substantially more non-stationarity than others; loss and throughput were considerably less stationary.

Sharing of congestion points. We analyze the degree of bottleneck sharing for destination clusters using local DNS server entries to identify live IP addresses in the same prefix clusters. These prefix clusters are in turn identified using BGP

data based on the scheme proposed by Krishnamurthy and Wang [13]. Previous work has also focused on identifying whether certain flows share the same point of congestion [21], [6], [11], [12]. However, they do not identify the point of congestion. We focus on explicitly identifying shared bottlenecks rather than implicit inferences.

Correlation between different Internet path properties. Our work investigates the correlation between different path properties: bottleneck location, loss location and queueing delay values. To our knowledge, the correlation of the location of different path properties has not been studied. Previous work has focused on end-to-end path properties [23]. Moon *et al.* [19] discovered a periodic phenomena in the correlation between delay and loss. They conjecture that the cause is due to the synchronization effect of TCP reacting to shared loss events. Paxson [20] found that packet reordering is correlated with routing fluctuation.

Correlation with router and link properties. Our work correlates choke links with the related router and link properties. Similar work has been done in [5] and [3]. Choi *et al.* [5] has corroborated their point-to-point delay measurement to fiber maps and router configuration information. Agarwal *et al.* found little correlation between router CPU utilization and BGP updates [3].

VIII. Conclusion

In this paper, we present a measurement study characterizing network bottlenecks. We look at four Internet bottleneck properties: the persistence of bottleneck location, the extent of bottleneck sharing among destination clusters, the correlation with link loss and delay, and the relationship with router and link properties, including router CPU and memory load, the link capacity and traffic load. We find that 20% – 30% of the source-destination pair in our data set have perfect bottleneck persistence, and less than 10% of the IP addresses in the cluster share a bottleneck more than half of the time. We also observe that 60% of the bottlenecks on lossy paths can be correlated with a loss point no more than 2 hop away. The bottlenecks can be clearly correlated with link load, while there is no strong relationship with link capacity and the router CPU and memory load.

Acknowledgments

Ningning Hu and Peter Steenkiste were in part supported by the NSF under award number CCR-0205266.

References

[1] Tulip Manual. `www.cs.washington.edu/research/networking/tulip/bits/tulip-man.html`.

[2] University of Oregon Route Views Project. http://www.routeviews.org.

[3] S. Agarwal, C.-N. Chuah, S. Bhattacharyya, and C. Diot. Impact of BGP dynamics on router CPU utilization. In *Passive and Active Measurement Workshop*, France, 2004.

[4] H. Balakrishnan, S. Seshan, M. Stemm, and R. Katz. Analyzing Stability in Wide-Area Network Performance. In *Proc. ACM SIGMETRICS*, June 1997.

[5] B.-Y. Choi, S. Moon, Z.-L. Zhang, K. Papagiannaki, and C. Diot. Analysis of point-to-point packet delay in an operational network. In *Proc. IEEE INFOCOM*, March 2004.

[6] K. Harfoush, A. Bestavros, and J. Byers. Robust identification of shared losses using end-to-end unicast probes. In *Proc. International Conference on Network Protocols*, November 2000.

[7] N. Hohn, D. Veitch, K. Papagiannaki, and C. Diot. Bridging Router Performance and Queuing Theory. In *Proc. ACM SIGMETRICS*, June 2004.

[8] N. Hu, L. E. Li, Z. M. Mao, P. Steenkiste, and J. Wang. Locating internet bottlenecks: Algorithms, measurements, and implications. In *Proc. ACM SIGCOMM*, August 2004.

[9] N. Hu and P. Steenkiste. Evaluation and characterization of available bandwidth probing techniques. *IEEE JSAC Special Issue in Internet and WWW Measurement, Mapping, and Modeling*, 21(6), August 2003.

[10] M. Jain and C. Dovrolis. End-to-end available bandwidth: Measurement methodology, dynamics, and relation with TCP throughput. In *Proc. ACM SIGCOMM*, August 2002.

[11] D. Katabi and C. Blake. Inferring congestion sharing and path characteristics from packet interarrival times. Technical report, Lab for Computer Science, MIT, Cambridge, MA 02139, 2001.

[12] M. S. Kim, T. Kim, Y. Shin, S. S. Lam, and E. J. Powers. A wavelet-based approach to detect shared congestion. In *Proc. ACM SIGCOMM*, August 2004.

[13] B. Krishnamurthy and J. Wang. On network-aware clustering of Web clients. In *Proc. ACM SIGCOMM*, August/September 2000.

[14] C. Labovitz, A. Ahuja, and F. Jahanian. Experimental Study of Internet Stability and Wide-Area Network Failures. In *Proc. International Symposium on Fault-Tolerant Computing*, June 1999.

[15] C. Labovitz, R. Malan, and F. Jahanian. Internet routing stability. *IEEE/ACM Trans. Networking*, 6(5):515–528, October 1998.

[16] C. Labovitz, R. Malan, and F. Jahanian. Origins of pathological Internet routing instability. In *Proc. IEEE INFOCOM*, March 1999.

[17] R. Mahajan, N. Spring, D. Wetherall, and T. Anderson. User-level internet path diagnosis. In *Proc. SOSP*, October 2003.

[18] Z. M. Mao, J. Rexford, J. Wang, and R. Katz. Towards an Accurate AS-level Traceroute Tool. In *Proc. ACM SIGCOMM*, September 2003.

[19] S. B. Moon, J. Kurose, and D. Towsley. Correlation of packet delay and loss in the internet. Technical report, Department of Computer Science, University of Massachusetts, Amherst, MA 01003, 1998.

[20] V. Paxson. *Measurements and Analysis of End-to-End Internet Dynamics*. PhD thesis, U.C. Berkeley, May 1996.

[21] D. Rubenstein, J. Kurose, and D. Towsley. Detecting shared congestion of flows via end-to-end measurement. In *Proceedings of the 2000 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 145–155, 2000.

[22] N. Spring, R. Mahajan, and D. Wetherall. Measuring ISP topologies with rocketfuel. In *Proc. ACM SIGCOMM*, August 2002.

[23] M. Tsuru, N. Ryoki, and Y. Oie. On the correlation of multiple path properties simultaneously measured on the Internet. In *Proceedings of SPIE Performance and Control of Next Generation Communications Networks*, volume 5244, pages 206–213, September 2003.

[24] Y. Zhang, N. Duffield, V. Paxson, and S. Shenker. On the constancy of internet path properties. In *ACM SIGCOMM Internet Measurement Workshop*, San Francisco, CA, November 2001.

[25] Y. Zhang, V. Paxson, and S. Shenker. The stationarity of internet path properties: Routing, loss, and throughput. Technical report, May 2000.