Analysis of Point-To-Point Packet Delay in an Operational Network ⋆

Baek-Young Choi [a,*], Sue Moon [b], Zhi-Li Zhang [c], Konstantina Papagiannaki [d], Christophe Diot [e]

[a] *University of Missouri, Kansas City, MO, USA*
[b] *Korea Advanced Institute of Science and Technology, Daejeon, Korea*
[c] *University of Minnesota, Twin Cities, MN, USA*
[d] *Intel Research, Cambridge, UK*
[e] *Thomson Research, Paris, France*

Abstract

In this paper we perform a detailed analysis of point-to-point packet delay in an operational tier-1 network. The point-to-point delay is the time experienced by a packet from an ingress to an egress point in an ISP, and it provides the most basic information regarding the delay performance of the ISP's network. Using packet traces captured in the operational network, we obtain precise point-to-point packet delay measurements and analyze the various factors affecting them. Through a simple, step-by-step, systematic methodology and careful data analysis, we identify the major network factors that contribute to point-to-point packet delay and characterize their effect on the network delay performance. Our findings are: 1) delay distributions vary greatly in shape, depending on the path and link utilization; 2) after constant factors dependent only on the path and packet size are removed, the 99th percentile variable delay remains under 1 ms over several hops and under link utilization below 90% on a bottleneck; 3) a very small number of packets experience very large delay in short bursts.

Key words: delay, ECMP, operational network

1. Introduction

In this paper we carry out a large-scale delay measurement study using packet traces captured from an operational tier-1 network. We focus on the so-called *point-to-point* (or, router-to-router) delay – the time between a packet entering a router in one PoP (the ingress point) and its leaving a router in another PoP (the egress point). Previously, Papagiannaki et al. [7] have measured and analyzed single-hop delay in a backbone network. Our work is the extension of [7] to the multiple hops case. The point-to-point delay measures the one-way delay experienced by packets from an ingress point to an egress point across an ISP's network and provides the most basic information regarding the delay performance of the ISP's network [6]. The objective of our study is two-fold: 1) to analyze and characterize the point-to-point packet delays in an operational network; and 2) to understand the various factors that contribute to point-to-point delays and examining the effect they have on the network delay performance.

Delay between two end-users (or points) has been studied extensively for its variation, path symmetry, queueing delay, correlation with loss, and more in [1, 9]. Because of its direct implication on delay-sensitive applications, such as VoIP (Voice over

⋆ An earlier version of this paper was presented in the proceedings of Infocom'04.

* Corresponding author. tel.:+1-816-235-2750; fax: +1-816-235-5159.
Email address: choiby@umkc.edu (Baek-Young Choi).

IP) and video streaming, and user-perceived performance of web downloads, there are continuing efforts on measuring, monitoring, and analyzing the end-to-end delay. Since the end-to-end delay is over several hops and may reflect route changes, it is not easy to pinpoint a cause of significant change, when we observe one. Point-to-point delay in an ISP is a building block of end-to-end delay, and understanding the main factors of point-to-point delay will add insight to the end-to-end delay.

The study of point-to-point packet delay poses several challenges.

- In order to understand the evolution of point-to-point delay, packet measurements need to span over a long period of time (e.g. hours).
- Data should be collected simultaneously from at least two points within an ISP, and clocks should be globally synchronized to compute one-way delay.
- Routing information is needed to keep track of route changes, if any. Other supplementary data, such as fiber maps and router configuration information, is needed to address path-specific concerns.

In our monitoring infrastructure, we have addressed all of the above points [4], and we believe this is a first study that focuses on point-to-point delay within an ISP.

We use packet traces captured in an operational tier-1 network. Packets are passively monitored [4] at multiple points (routers) in the network with GPS-synchronized timestamps. Captured packets between a pair of any two points are the entire traffic between the two, not from active probes or from a certain application. Using these precise delay measurement data, as well as SNMP (Simple Network Management Protocol) [11] link utilization data, router configuration information, routing updates, and fiber maps, we perform a careful analysis of the point-to-point delay distributions and develop a systematic methodology to identify and characterize the major network factors that affect the point-to-point packet delays.

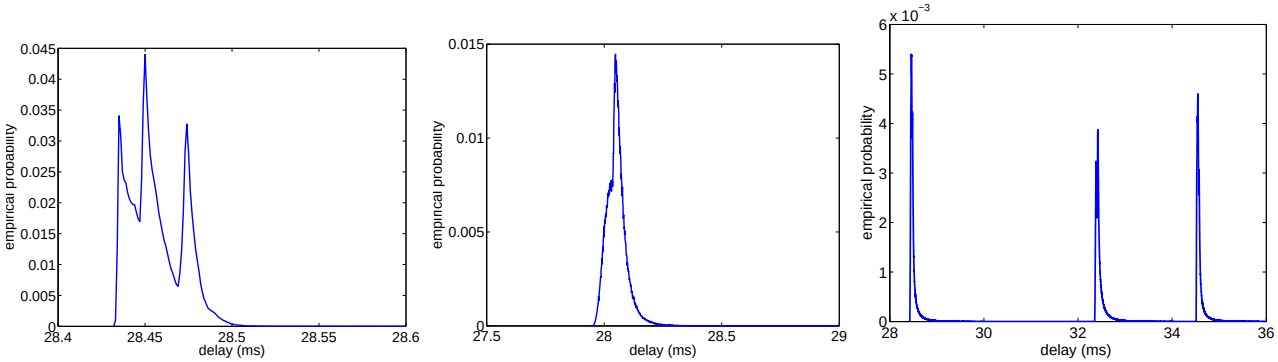
Our observations and findings are the following. First, the point-to-point packet delay distributions in general exhibit drastically different shapes, often with multiple modes, that cannot be characterized by a single commonly known mathematical distribution (e.g., normal or heavy-tailed distribution). There are many factors that contribute to these different shapes and modes. One major factor is the equal-cost multi-path (ECMP) routing [12] commonly employed in operational networks. It introduces multiple modes in point-to-point delay distributions. Another major factor is the packet size distribution, which has a more discernible impact on point-to-point packet delay distributions when the network utilization is relatively low. By identifying and isolating these factors through a systematic methodology with careful data analysis, we then focus on the variable delay components and investigate the role of link utilization in influencing the point-to-point packet delays. We find that when there is no bottleneck link on a path with utilization over 90%, the 99th percentile variable delay is less than 1 ms. When a link on the path has link utilization above 90%, the weight of the variable delay distribution shifts and the 99th percentile reaches over 1 ms. Even when the link utilization is below 90%, a small number of packets experiences delay one order of magnitude larger than the 99th percentile and affects the tail of the distribution beyond the 99th percentile.

In summary, the contribution of this paper first lies in the detailed analysis and characterization of the point-to-point packet delays and their major components in an operational network. We also develop a systematic methodology to identify, isolate, and study the main contributing factors. Understanding when and how they affect the point-to-point packet delays in an operational network is important both theoretically and in practice: such an understanding will not only help network engineers and operators to better engineer, operate, and provision their networks, it also provides valuable insight as to the important performance issues in operational networks that network researchers should pay attention to. In particular, our study sheds light on the necessity and importance of devising more representative and relevant measurement mechanisms for network delay performance and link utilization monitoring.

The remainder of the paper is structured as follows. Section 2 lays out the factors that affect the point-to-point delay and describes our packet measurement methodology and other data, such as SNMP statistics and router configuration information. The main discussion begins in Section 3 with general observations on point-to-point delay. Then in Section 4 we isolate constant factors that are fixed on the path and the packet size. In Section 5, we focus only on the variable part of the delay that is due to cross traffic. In Section 6, we investigate further SNMP data for the cause of long delays. We

Table 1
Summary of Matched Traces: Delays are in milliseconds.

Data Set	Link Speed (From)	Link Speed (To)	Duration	Packets	Min	Mean	Median	99th	Max.
1	OC-48	OC-12	16h 24m	1,349,187	28.432	28.458	28.453	28.494	85.230
2	OC-12	OC-12	3h	498,865	27.949	28.057	28.051	28.199	55.595
3	OC-12	OC-48	5h 21m	4,295,781	28.424	31.826	32.422	34.894	100.580



(a) Data Set 1 (b) Data Set 2 (c) Data Set 3

Fig. 1. Point-to-point packet delay distributions

summarize our findings and discuss their implications on current monitoring practice in Section 7.

2. Point-to-Point Delay and Measurement Methodology

We define *point-to-pont* packet delay as the time between a packet entering a router in one PoP (the ingress point) and its leaving a router in another PoP (the egress point). Theoretically speaking, we can decompose the point-to-point packet delay into four components: propagation delay, transmission delay, nodal processing delay and queueing delay. Propagation delay is determined by physical characteristics of the path a packet traverses, such as the physical medium and its length. Transmission delay is a function of the link capacities along the path, as well as the packet size. Nodal processing delay is the time to examine the packet header and determine the route for the packet. It also includes checksum calculation time and the transfer time from an input to an output port. On today's high-speed routers it is typically less than 30 μs [7]. Queueing delay depends on the traffic load along the path, and thus varies over time. In practice, many other factors can contribute to the delay packets experience in an operational network. First, network routing may change over time, hence the path between an ingress point and an egress point may not be fixed. Furthermore, in today's high-speed backbone networks, equal-cost multi-path (ECMP) routing is commonly employed for traffic engineering and load balancing. Hence packets going through the same ingress-egress point pair may take different paths with differing path characteristics such as propagation delay, link capacities, and traffic load. These factors can introduce significant variations in point-to-point delay measurement. We will refer to factors that depend solely on path characteristics as well as packet sizes as constant factors, since these factors have a fixed effect on point-to-point delays experienced by packets of the same size that traverse exactly the same path.

Queueing delay introduces a *variable* component to the point-to-point delays experienced by packets, as it depends on the traffic load (or cross traffic) along the path, which varies at different links and with time. In addition to traffic load, sometimes anomalous behaviors of individual routers or the network can also introduce other unpredictable factors that affect point-to-point delays packets experience. For example, in an earlier study [7], authors discovered that packets may occasionally experience very large delays. Such large delays can be caused by a router performing other functions such as routing and forwarding table updates, etc. In addition,

during the routing protocol convergence, transient forwarding loops [5] may occur, and packets caught in such loops will suffer unusual long delays. These very large delays are obviously not representative of the typical delay performance of a network, and thus should be considered as outliers in delay measurement.

With a basic understanding of the various factors affecting point-to-point packet delays in an operational network, we now proceed to describe the delay measurement setting and methodology employed in our study. The packet traces we use in this study are from the Sprint IPMON project [4]. On the Sprint IP backbone network, about 60 monitoring systems are deployed at 4 PoPs (Point-of-Presences), each with many access and backbone routers. Using optical splitters, the monitoring systems capture the first 44 bytes of all IP packets and timestamp each of them. As the monitoring systems use the Global Positioning System (GPS) for their clock synchronization, the error in timestamp accuracy is bounded to less than 5 μs.

To obtain packet delays between two points (i.e., between two routers in two different PoPs), we first identify those packets that traverse those two points of measurement. We use hashing to match packets efficiently. Only 30 bytes out of the first 44 bytes of a packet carry distinguishing information and are used in hashing. IP header fields, such as the version, TTL (Time-To-Live), and ToS (Type of Service), are not used. For more detail about the packet matching, refer to [7].

While matching packets, we occasionally find duplicate packets in our traces. They are likely due to unnecessary link-level retransmission or routing loops, and have been reported in [5, 7, 8]. Since duplicate packets have the identical 30 bytes we use in hashing, we cannot always disambiguate packets in the corresponding traces and do not use them in our analysis. In all pairs of the traces, we observe that duplicate packets are less than 0.05% of the total number of packets.

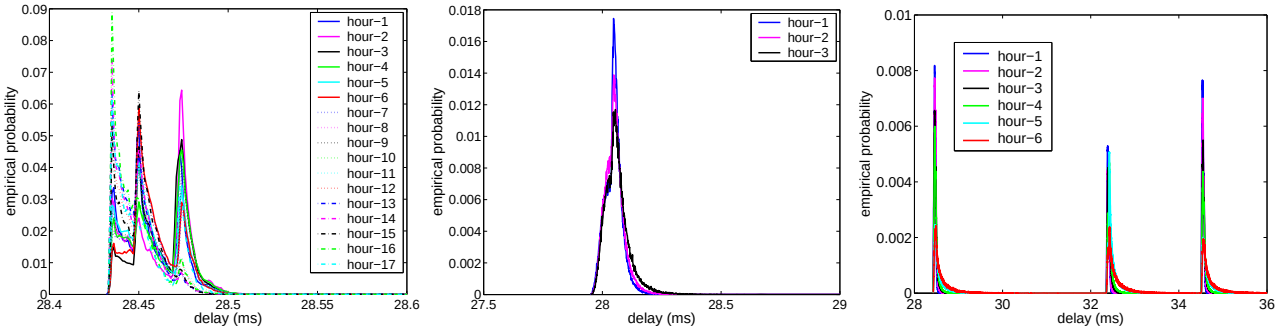
We have analyzed weeks worth of concurrent traces dispersed over a few years. For this work, we select traces from 2 dates: August 6th, 2002, and November 21st, 2002. The main criterion used in trace selection is the number of packets common in a pair of traces from router links. Not all pairs of traces have many packets in common. Traffic entering a network at one ingress point is likely to split and exit through many egress points. Thus the number of packets observed on a pair of monitored links depends on the actual amount of traffic that flows from one to the other. Some pairs of traces have no packet in common. As our goal is to analyze packet delay, we choose to use pairs of traces with most matches in our analysis.

In conducting our delay measurement study, we have analyzed all matched trace pairs from August 6th, 2002, and November 21st, 2002. In this paper, however, we choose only three pairs as representative examples to illustrate our observations and findings. The delay statistics of the three trace sets we use in this paper are shown in Table 1. In all three sets, the source traces are from the West Coast and the destination traces are from the East Coast of the United States. Even though we could not have inter-continental link traces for the analysis, the selected data sets illustrate the U.S. transcontinental delays. The path between a source and a destination consists of more than 2 hops, and thus the delay reported in this work is over multiple hops. The duration of the matched traces varies from around 3 hours to more than 16 hours.

In addition to the packet traces, we also use other network data such as SNMP statistics on link and CPU utilization, router configuration files, and fiber maps in our analysis and identification of various network factors that affect the measured point-to-point packet delays. Using the router configuration files and routing information, we obtain the information about the paths, the associated IP links, and router interfaces that packets traverse in the point-to-point packet delay measurements. The fiber map provides us with the further information about the fiber links and estimated propogation delay of the paths. SNMP data, which report the link load and router CPU utilization averaged over five-minute intervals, are also collected on every link along each path, and are used to correlate the point-to-point packet delay measurements with 5-minute average link utilization.

3. General Observations

We begin this section with general observations on point-to-point packet delay distributions obtained from the three trace sets in Table 1. First, we note from Table 1 that the minimum delays from all three delay distributions are about 28 ms, which reflect the transcontinental delay of the U.S. Other statistics



(a) Data Set 1 (b) Data Set 2 (c) Data Set 3

Fig. 2. Hourly point-to-point packet delay distributions

such as mean, median, 99th percentile [1] and maximum delays show more variations. However, when we examine the delay distributions of the three sets, the difference between the traces are striking. We use 1 μs bins to plot the empirical density distribution functions of the packet delay. The bin size is chosen to be small enough to exhibit important modes. In order to show the raw data we did not use a kernel based or average shifted histogram, though they may be used to smooth erratic modes that do not relate to any underlying physical effect. The resulting point-to-point delay distributions in the entire duration of the traces are shown in Figure 1. Clearly, the shapes of the three (empirical) delay distributions are starkly different.

Figure 1(a) exhibits three peaks that are apart from each other only in tens of microseconds. Figure 1(b) has only one peak and most of the distribution lies between 27.9 and 28.5 ms. Figure 1(c) is very different from the other two: it has three unique peaks, which are apart from each other by 2 and 4 ms, respectively. Here we point out that the x-axes of the three plots are in different ranges: they are chosen to include the 99th quantile point, but not the maximum delay point. Though the difference between the minimum and the 99th percentile delay is less than 1 ms in Data Sets 1 and 2, and 6.5 ms in Data Set 3, the maximum delay is significantly larger than the 99th percentile in all three sets. As the number of packets with such extreme delay is very small, they represent very rare events in the network.

We take a more detailed look at the delay distributions and how they change over time. We divide each data set into segments of an hour long and plot the hourly point-to-point delay distributions and see whether they look significantly different from that of the overall data sets. Figure 2 [2] shows the hourly delay distributions overlaid on top of each other for each of the three data sets. Clearly, the basic shapes of the three distributions remain the same in these hourly plots. In particular, the peaks or modes of the distributions remain at the same locations; there are three peaks within a very short range in Figure 2(a), a single peak in Figure 2(b), and three peaks of much distance between them in Figure 2(c). However, the bulk as well as the tail of the hourly delay distributions show discernible variations in all three trace pairs: some hourly delay distributions have shorter peaks and fatter tails.

What contributes to the differences in the point-to-point packet delay distributions in an operational network? More specifically, what network factors cause the differing numbers of peaks or modes in the delay distributions? What affect the bulk and (or) tail of the delay distributions? These are the main questions we set out to answer in this paper. We develop a systematic methodology to identify and characterize the various network factors that contribute to the point-to-point packet delays through careful data analysis, as well as using additional information such as router configuration information, routing and SNMP data. In the following sections, we will investigate these factors one by one methodically.

[1] Let $F(x)$ be a cumulative distribution function of a random variable x. $q = F^{-}(0.99)$ is the 99th percentile of the distribution of x.

[2] The multiple lines may not be easily distinguished. However it should not trouble the reading of the paper, as precise differentiation is not necessary.

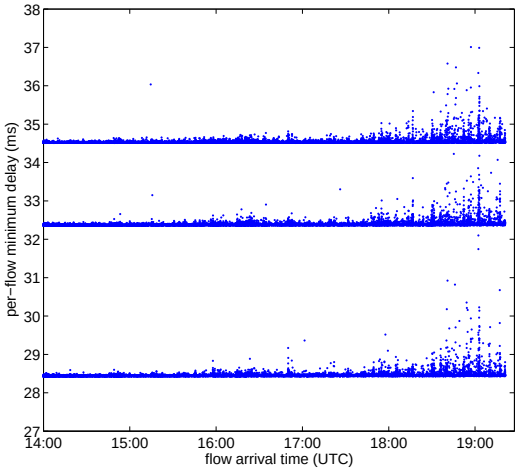


Fig. 3. Minimum flow delay vs. flow arrival time of Data Set 3

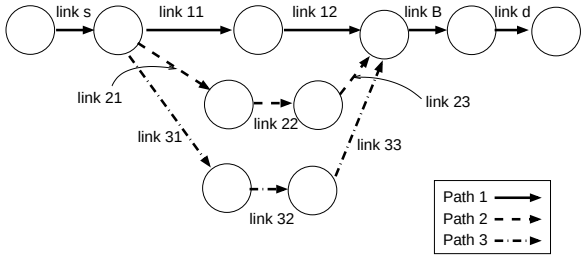
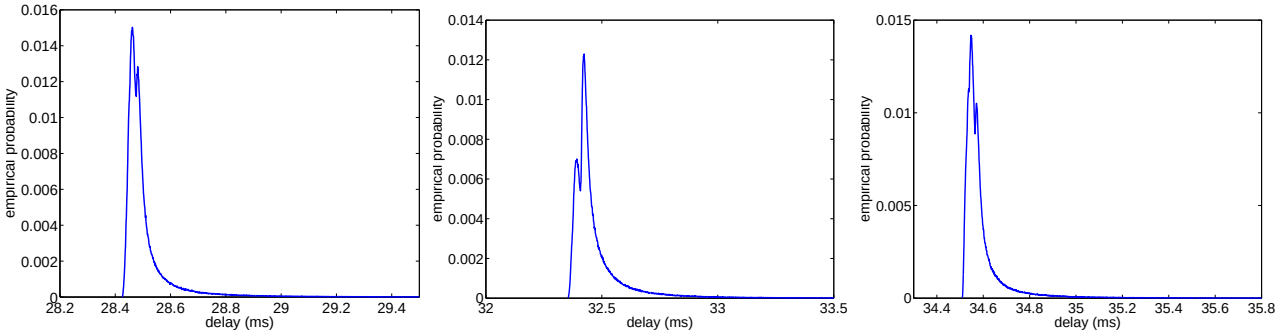


Fig. 4. Path connectivity



(a) Path 1 (b) Path 2 (c) Path 3

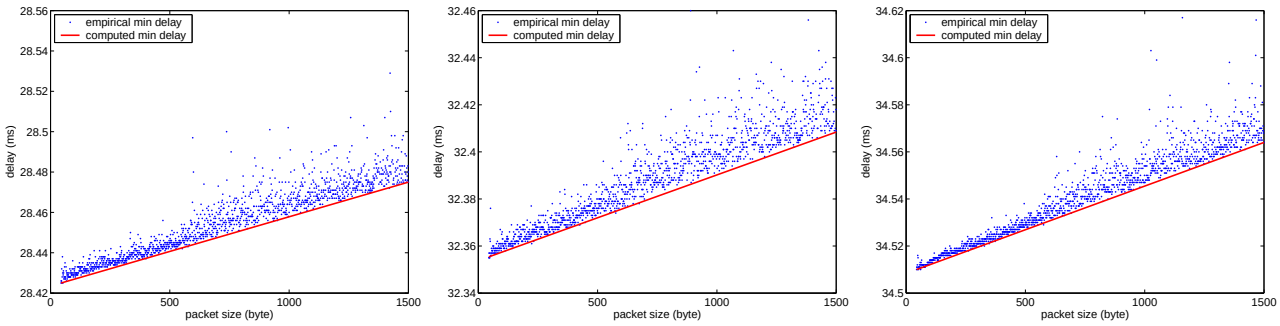
Fig. 5. Delay distributions

4. Identification of Constant Factors

In this section we characterize and isolate the constant network factors that have a fixed effect on point-to-point delays experienced by packets of the same size traversing the same *physical* (i.e., fiber) path. In particular, we identify and analyze the constant network factors that contribute to the *modes* in the point-to-point packet delay distributions shown in the previous section. We suspect that the modes of the delay distribution that are spaced with relatively large distance (1 ms or more), such as in Figure 1(c), are most likely caused by the equal-cost multi-path (ECMP) routing, commonly employed in today's operational networks. On the other hand, the modes that are more closely spaced (within 10s or 100s of microseconds), such as in Figure 1(a), are probably due to the effect of various packet sizes that incur different transmission delay. The latter factor, for example, has been shown to have an impact on single-hop packet delays [7]. In the following we develop a systematic methodology to identify and isolate such effects.

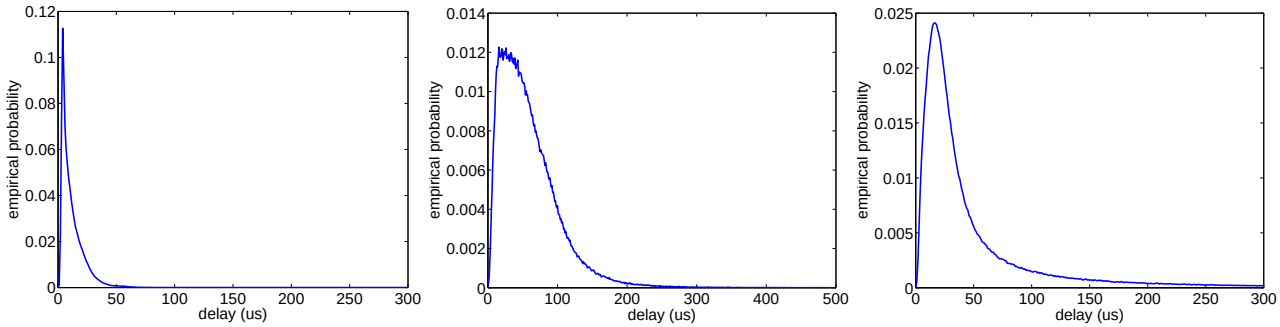
4.1. *Equal-Cost Multi-Path Effect*

As mentioned earlier, equal-cost multi-path (ECMP) routing is commonly used in today's tier-1 operational networks for load balancing and traffic engineering. Here equal cost refers to the sum of the weights across the shortest paths used by intra-domain routing protocols such as ISIS. These weights are not necessarily related to the physical properties of a path such as its propagation delay. In fact, sometimes paths that follow separate fiber circuits are preferred for fault tolerance. Because of the differing characteristics of these physical paths, their propagation and transmission delay may also



(a) Path 1 (b) Path 2 (c) Path 3

Fig. 6. Minimum packet delay vs. packet size



(a) Data Set 1 (b) Data Set 2 (c) Path 1 of Data Set 3

Fig. 7. Delay distributions after removing $d^{fixed}(p)$

be different. In using ECMP routing, routers (e.g., Cisco routers in our study) randomly split traffic using a hash function that takes the source IP address, the destination IP address, and the router ID as input to determine the outgoing link for each packet [3]. Therefore packets with the same source and destination IP addresses always follow the same path.

To identify the effect of ECMP routing on point-to-point packet delays, we employ the following method. We first define a *(two-tuple) flow* to be a set of packets with the same source and destination IP addresses, and group packets into flows. We then compute the *minimum* packet delay for each flow. The idea is that by considering the minimum packet delay of a flow, we attempt to minimize the variable delay components such as queueing and retain mostly the fixed delay components. If the minimum delays of two flows differ significantly, they are more likely to follow two different paths. In Figure 3 we plot the minimum delay of each flow by the arrival time of the first packet in the flow for Data Set 3. The plot demonstrates the presence of three different paths, each corresponding to one peak in the delay distribution of Figure 1(c). We number the path with the smallest delay (28.4 ms), Path 1, the next (32.4 ms), Path 2, and the last with the largest delay (34.5 ms), Path 3.

We use other network data to corroborate our finding. Using the router configuration and fiber path information, we identify the three paths and the exact links the packets in Data Set 3 traverse. The topology map is in Figure 4. The source is marked link s, the destination, link d, and the three paths share one common link (denoted link B), one hop before the destination. In addition, using the fiber map, we also verify that the fiber propagation delay (28 ms, 32 ms, and 34 ms) matches the minimum delay of each of the paths.

With the knowledge of the three paths, the next task is to isolate the effect of ECMP, namely, to clas-

[3] By having the router ID as input to the hash function, each router makes a traffic splitting decision independent of upstream routers.

sify and separate the packets based on the path they take. From Figure 3, it is clear for most of the flows which path they take. However, near the end, minimum delays of some flows become close to the minimum of the next path, and it becomes hard to distinguish which path they take. Hence the minimum delay of a 2-tuple flow alone is not sufficient to pin down the path it takes during this period. To address this issue, we take advantage of other fields in the packet header. Every IP packet carries a 1-byte long TTL (Time-to-Live) in the header. Because the TTL value is decremented by one at every hop, the difference in the TTL values between two points of observation tells the number of hops between them. When we examine the TTL field of packets from the first hour of the trace (when the per-flow minimum delay can easily indicate the path a flow took), packets from those flows whose minimum delay is close to 28.4 ms have a TTL delta of 4. That is, Path 1 consists of four hops from the source to the destination. Other flows, whose minimum delay is above 32.4 ms, all have a TTL delta of 5.

Using the TTL information, we separate packets that follow Path 1 from Data Set 3. For the remaining packets, we classify those flows with delay less than 34.5 ms to Path 2, and the rest to Path 3. Because Paths 2 and 3 have the same TTL delta, near the end of the trace we cannot completely disambiguate the paths some flows take and have some packets misclassified. However, those flows whose minimum delay is high away from the closest path transit time, tend to consist of only a few (one or two) packets. Thus, the number of such packets is considered extremely small, compared to the total number of packets in the trace.

Figure 5 shows the delay distribution for each of the three paths. We see that the shape resembles more like that of Figure 1(b) with two barely discernible modes. In fact, the modes in Figure 1(a) and Figure 5 are due to packet size, which is another constant factor of delay. We discuss the impact of packet size in detail in the next subsection.

4.2. *Minimum Path Transit Time*

With the ECMP effect removed from point-to-point packet delay, we now focus on characterizing the fixed delay component caused by such constant network factors as propagation delay, transmission delay, per-packet router processing time, etc. Theoretically speaking, given a packet of size p that traverses a path of h hops, each link of capacity C_i and propagation delay δ_i, the total propagation and transmission delay can be written as:

$$d^{fixed}(p) = \sum_{i=1}^{h}(p/C_i + \delta_i) = p\sum_{i=1}^{h} 1/C_i + \sum_{i=1}^{h}\delta_i.$$

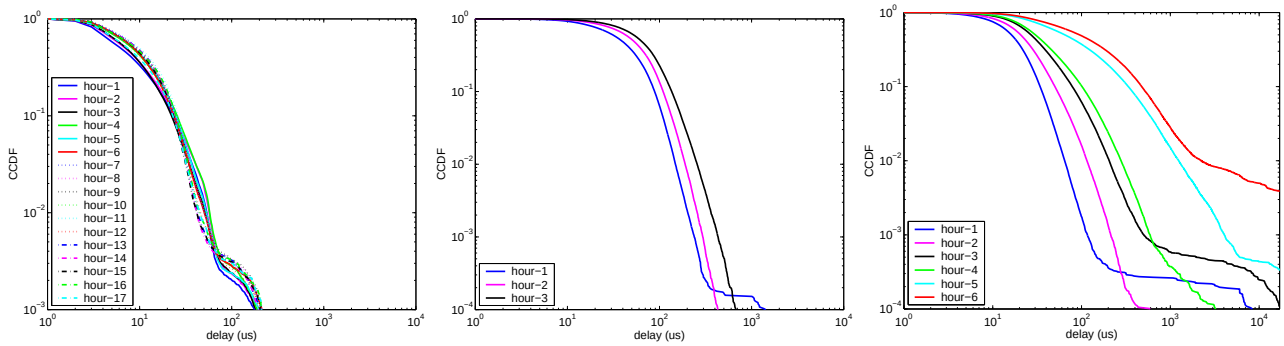
In other words, the fixed delay component (i.e., the total propagation and transmission delay) for a packet of size p is a *linear* (or precisely, an *affine*) function of its size,

$$d^{fixed}(p) = \alpha p + \beta \tag{1}$$

where $\alpha = \sum_{i=1}^{h} 1/C_i$ and $\beta = \sum_{i=1}^{h}\delta_i$.

In practice, however, in addition to link propogation and transmission delay, routers also introduce other factors such as per-packet router processing time, backplane transfer inside a router, and so forth that are (mostly) dependent on packet size. Hence in reality we believe that such a linear relation between the fixed delay component and packet size is still valid, albeit the parameters α and β will not be a simple function of the link capacities and the propogation delays as in (1). In fact, in analyzing single hop packet delay, the authors in [7] show that the minimum router transit time through a single router is linearly proportional to the packet size. Assuming that the same linear relation holds at all routers on the path, we believe the minimum time any packet experiences on a path of multiple hops to be also linearly proportional to the packet size. To validate this assumption, we check the minimum delay of packets of the same size for each path, and plot the minimum delay against the packet size. Figure 6 shows the corresponding plots for the three paths in Data Set 3 using packets from the first hour. As expected, there is an apprarent linear relation. We fit a line through linear regression on the packet sizes of 40 and 1500 bytes, since they are the most common sizes and their minimum delays are most likely to represent the real minimum on the path. This line yields a minimum fixed delay for a given packet size, and we refer to it as the *minimum path transit time* for the given packet size, denoted by $d^{fixed}(p)$. The $d^{fixed}(p)$ was the same with the packets from other hours. Using routing information, we confirmed that the routing was stable for the duration of the trace collection.

The parameters α and β derived from the plots in Figure 6 are listed in Table 2. Note in particular that Path 2 and Path 3 have the same α value, but different β values. This is consistent with the fact



(a) Data Set 1 (b) Data Set 2 (c) Path 1 of Data Set 3

Fig. 8. Hourly distribution of variable delay

that the two paths have exactly the same number of hops, same link transmission speed and same type of routers, but slightly different propogation delay along their respective paths due to the difference in the length of their fiber paths. Using packets from other hours in Data Set 3 we obtain almost identical α and β values for each path, again confirming the linear relation between the minimum path transit time and packet size. The same result holds for other data sets we have analyzed.

With the fixed delay component $d^{fixed}(p)$ identified, we can now substract it from the point-to-point delay of each packet to study the *variable delay component* (or simply, *variable delay*). Let d represent the point-to-point delay of a packet. The variable delay component of the packet, d^{var}, is given by $d^{var} := d - d^{fixed}(p)$. In the next section we investigate in detail the distributions of variable delays, $\{d^{var}\}$, experienced by packets, how they change over time, and what the major factors are that contribute to long variable delays.

In Figure 7 we plot the distribution of variable delay (i.e., after the fixed delay component has been removed) for Data Set 1, Data Set 2, and Path 1 of Data Set 3. The minor peaks we observe in Figures 1(a) and Figure 5 disappear, and now we only see uni-modal distributions in all the figures [4] .

[4] Due to space limitation, we do not include the variable delay distributions for Paths 2 and 3 of Data Set 3. They are similar to that of Path 1.

Table 2
Slope and y-intercept of minimum path transit time

Data Set	Path	y-intercept (β)	slope (α)
1		28.430931	0.00002671
2		27.947616	0.00005959
3	1	28.423489	0.00003434
3	2	32.353368	0.00003709
3	3	34.508368	0.00003709

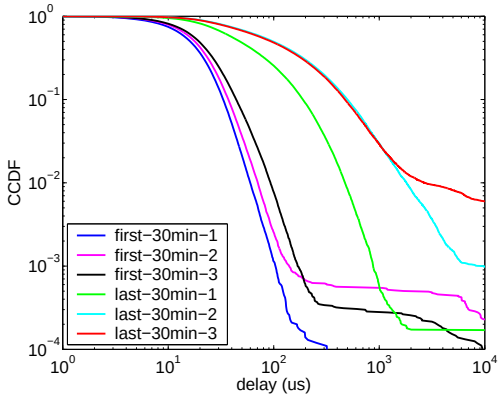
5. Analysis of Variable Components

5.1. *Variable Delay and Link Utilization*

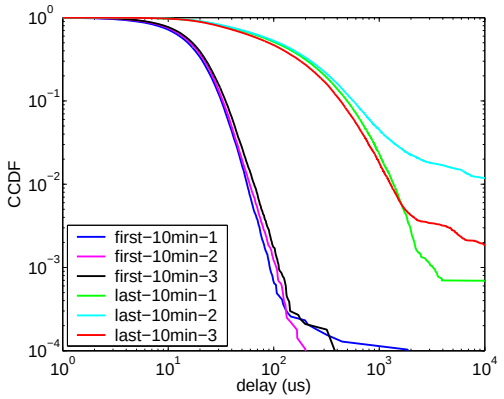
To understand how the distribution of variable delay changes over time, we plot the hourly distributions of the variable delay for Data Set 1, Data Set 2, and Path 1 of Data Set 3 in Figure 8. The hourly distributions are overlaid over each other for ease of comparison. Here we use the complimentary cumulative distribution function (CCDF), as it more clearly illustrates where the bulk of the distribution lies and how the tail tapers off.

From Figure 8(a), the hourly distributions of variable delay for Data Set 1 are nearly identical, signifying that the network condition did not change much throughout the entire duration of our measurement. Also the bulk of the distribution (e.g., 99.9th percentile) lies under 200 μs. Data Set 1 is from a path over 4 router hops, and thus less than 200 μs of variable delay is quite small. Hence packets traversing along this path experience very little queueing delay.

In Figure 8(b), the variable delay distributions display a slight shift from the first to the third hour, indicating that the network conditions during the 3 hour period have slightly changed. However the



(a) In 30-minute intervals



(b) In 10-minute intervals

Fig. 9. CCDF of $\{d^{var}\}$ of the first and last segments of Path 1 of Data Set 3.

bulk of the distributions (99.9th percentile) is still within hundreds of microseconds. Variable delay of less than 1 ms is generally not very significant, especially over multiple hops, and reflects well on the network performance an end-user should perceive.

The hourly distributions of Path 1 to Path 3 in Data Set 3, however, tell a very different story. The hourly plots shift significantly from hour to hour: the plots from the last two hours diverge from that of the first hour drastically, especially for the tail 10% of the distribution. The 99% delay in the first hour is 100 μs, while in the last hour it is at least an order of magnitude larger.

To examine the changes in the variable delay more closely, we zoom in and plot the delay distributions using smaller time intervals. Figure 9(a) show the distribution of variable delay in the first and last three 30-minute intervals for Path 1 of Data Set 1; Figure 9(b) show those from the first and last three 10-minute intervals. In the first 30 minutes, there is little change, and even within the first three 30 minutes, there is not much change in variable delay except for the tail beyond the probability of 10^{-3}. The 99.9th percentile of the distributions is still within a few 100 μs. However in the last three 30 minutes, in particular, the last 30 minutes, many packets experience much larger variable delay, causing the 99th percentile delay to reach 2 ms, and in case of a single 10-min interval, up to almost 10 ms. Though not presented here, the hourly plots from Paths 2 and 3 exhibit a similar shift toward the end of the trace: the variable delay increased significantly.

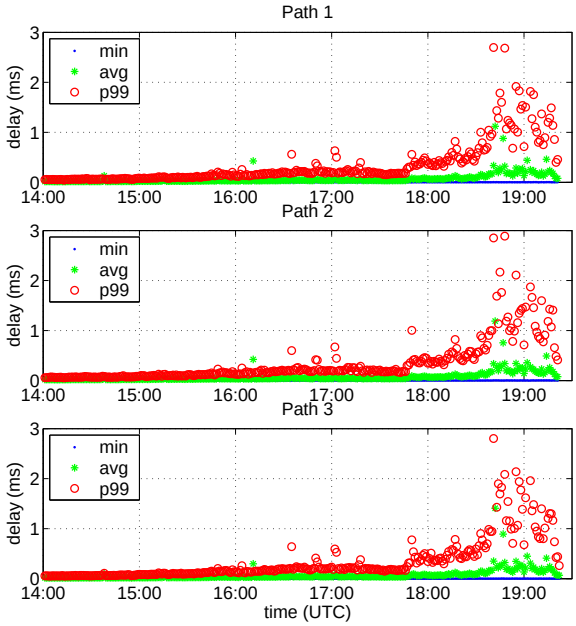


Fig. 10. Minimum, average, and 99th percentile of variable delay per 1-minute interval of Data Set 3

To investigate how the variable delay evolves throughout the entire trace, we compute the minimum, average, and 99th percentile of variable delay over every 1-minute interval. The corresponding time series plot is shown in Figure 10. The average and 99th percentile delays increase significantly near the last one hour of the trace. In the case of the 99th percentile delays, they often are above 1 ms, which seem to indicate a significant level of congestion during that time period. All three paths experience heightened level of congestion during about the same time period, which, in turn, makes us suspect the common link, link B, to be the bottleneck.

The link utilization on the path from the source to the destination should tell us if there was any link highly loaded or severely congested. For this purpose, we examine the SNMP statistics collected

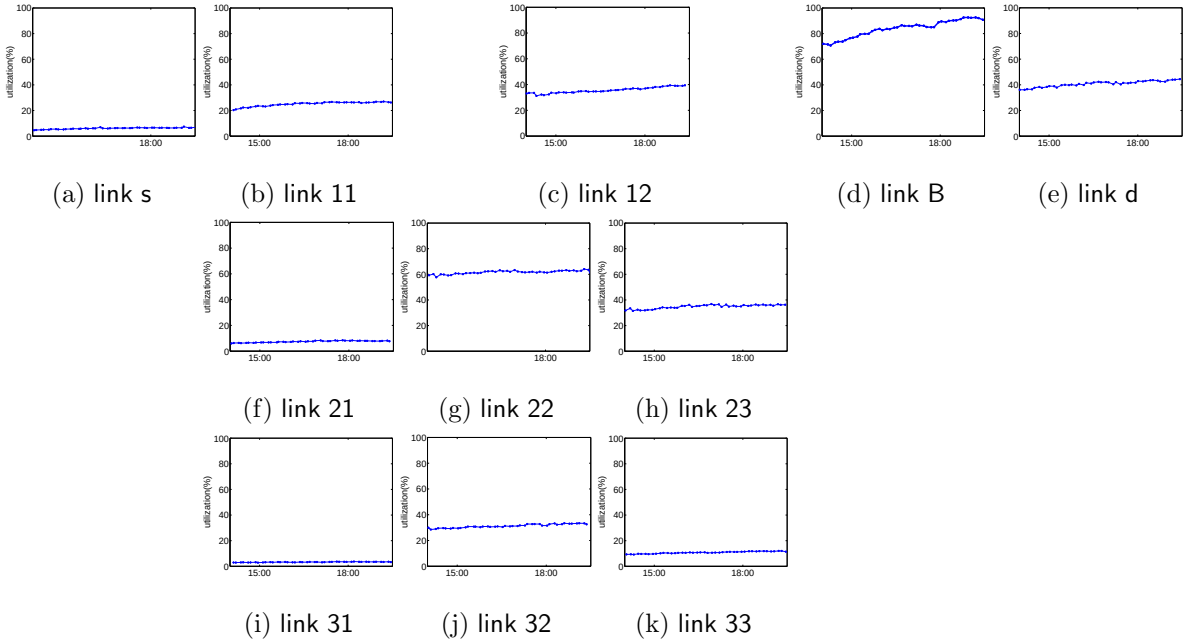


Fig. 11. SNMP statistics of link utilization in 5-minute intervals

on the links along the paths to validate what we observe in Figure 10. Figure 11 displays the link utilization on all links of the three paths. For ease of viewing, we place a plot of link utilization over time in a matching position to its corresponding link in the topology map shown in Figure 4. Link B had the highest link utilization of 70% even at the beginning of the trace, and the utilization level increased to 90% near the end of the trace. Clearly, link B was the bottleneck that caused the significant increase in delay.

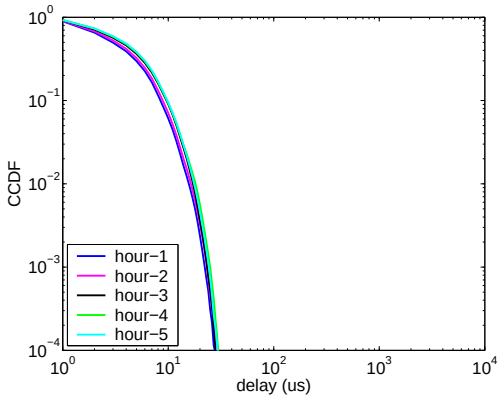
Another way to confirm that link B was truly the bottleneck is to compare the delay before and after the bottleneck point. We have an extra set of measurements from link12, and can calculate the delay from link s to link 12 and from link 12 to link d[5]. Unfortunately, the trace from link 12 is only 5 hours long, which does not include the last half an hour. Figure 12(a) presents the CCDF of variable delay from link s to link 12. All hourly plots are overlaid on top of each other, demonstrating that the network condition on the partial path had not changed much and the packets experienced almost no variable delay (less than 30 μs for 99.99% of the packets). Figure 12(b) shows hourly distributions of variable delay from link 12 to link d. They closely match the shape of Figure 8(c). We conclude that the high utilization on link B is the deciding factor that increased variable delay on the path from link s to link d.

[5] We do not have measurements from any other intermediate links on Paths 2 or 3.

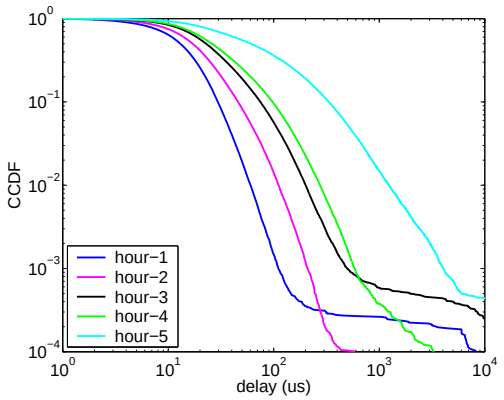
5.2. *Analysis of Peaks in Variable Delay*

In the previous section, we identify the high utilization on the shared link, link B, as the cause of large variable delay on all three paths of Data Set 3, thus the shift in the bulk of the delay distributions in the later hours. Another interesting phenomenon apparent in Figure 8(c) (also in Figure 9) is that the very tail in several of the hourly delay distributions flattens out abruptly and then tapers off. For example, the very tail (at the probability around $0.5x10^{-3}$) in the hour-1 delay distribution flattens out and reaches a very large delay of almost 10 ms, even though the average link utilization is only approximately 70%. What causes such a phenomenon? We address this question by analyzing the packets that experience the very large variables.

To examine when and how such very large delay occurs, we set the threshold of very large delay at 1 ms, and consider only packets with a variable delay greater than 1 ms. Figure 13 presents a time series for the variable delay of these packets vs. the time they arrive at the destination link d. We see that most of these packets appear clustered, and form six conspicuous peaks, labeled P-1 to P-6, respectively.



(a) From link s to link 12



(b) From link 12 to link d

Fig. 12. Empirical CCDF of $\{d^{var}\}$

The maximum delay in each peak exceeds 8 ms, and, in the case of P-5, reaches up to 100 ms. All the peaks last for a very short period of time, in contrast to the duration of 5.5 hours of our measurement. In search of the causes of the peaks, one possible explanation is to attribute them to some random or uncommon events caused, e.g., by router idiosyncrasies [7]. For instance, spiky behavior in end-to-end packet delay has been previously reported by Ramjee et al. [10]. In their measurements, a delay spike has no ascending slope, but instantly reaches the maximum value and then gradually decreases over time. Claffy et al. observe similar behavior in round-trip time delay and attribute it to Internet hardware problems associated with prefix caching on routers [3]. Before we dismiss these peaks as aberrant behavior, we zoom in at finer time scale to examine the details of these peaks.

In Figure 14, the zoomed-in pictures of the peaks (in the time scale seconds) reveal an interesting behavior: instead of an abrupt rise at the beginning of a

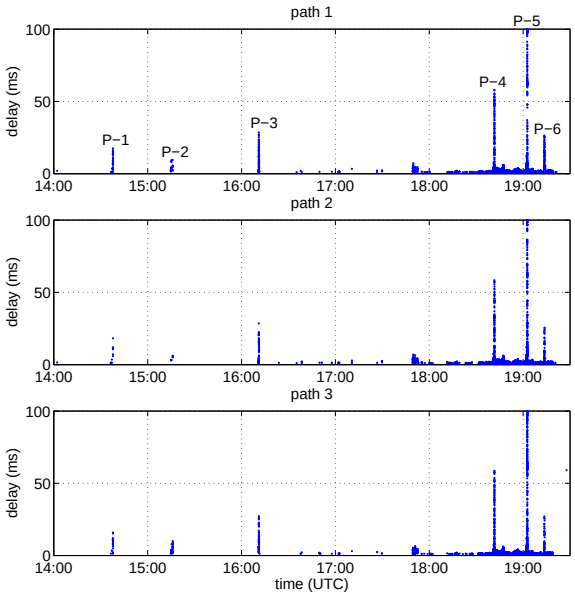


Fig. 13. $\{d^{var}\}$ above 1 ms from Data Set 3.

peak, the delay gradually increases, reaches a peak, and then decreases. It resembles more of the behavior of a queue building up gradually and eventually being drained. In P-2, it is hard to see the evolution of a peak. P-2 is the shortest, and the dataset may not contain sufficient packets during the P-2 peak formation period to sample the queue building up and draining process. In other peaks, the ascending slope is in general steeper than the descending one. We speculate that the magnitude of the increasing slope is related to the throughput of the aggregate flow that overloads the bottleneck. The decreasing slope is probably related to the output capacity of the bottleneck.

The most striking observation is from P-5. The delay reaches a plateau at 100 ms and remains at 100 ms for a little longer than 4 seconds. Delays of 100 ms on an OC-48 link translate to 30 MB of buffer space. We do not know the exact buffer size kept at the output queue for link B. Nor do we have packet-level measurements from link B or of other incoming traffic to link B, other than from link 12. At this point we do not have sufficient information to fully understand the underlying cause of P-5. We can only speculate that the output buffer reached its full capacity and thus the variable delay did not increase any more.

To see if these peaks occurred on the path segment before the bottleneck, link B, we zoom into the same time period as in Figure 14 at link 12. We see no

variable delay above 1 ms between link s and link 12. When examining the same time period on the path segment between link 12 and link d, we see peaks very similar to those in Figure 13. Figure 15 shows the zoomed-in picture corresponding to Figure 14. Note that the figure contains only four time series plots, this is because the trace collected on link 12 is shorter than the traces collected on link s and link d, as noted earlier. It is only 5 hours long, and does not cover P-5 or P-6. The four plots in Figure 15 match the first four in Figure 14 in both height and duration. This suggests that the delay peaks are caused by sudden traffic bursts on the bottleneck link B.

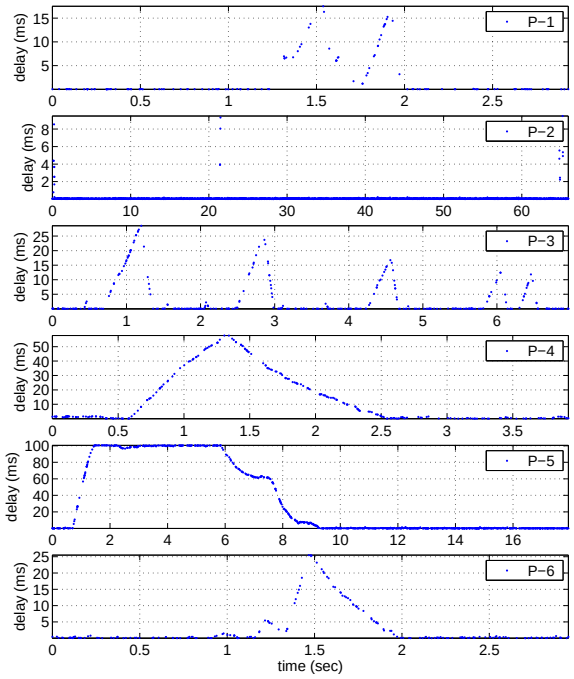


Fig. 14. $\{d^{var}\}$ of peaks, P-1 to P-6 on Path 1 of Data Set 3.

We summarize our findings in this section as follows. The variable delay distribution has two parts: the first part represents the bulk of the distribution (99th percentile). When the link utilization on a bottleneck link is below 90% on 5 minute average, the 99th percentile of the hourly delay distributions remains below 1 ms. Once the bottleneck link reaches utilization levels above 90%, the variable delay shows a significant increase overall, and the 99th percentile reaches a few milliseconds. The second part is about the very tail of the distribution. Even when the link utilization is relatively low (below 90% on 5 minute average), sometimes a small number of packets may experience delay an order of

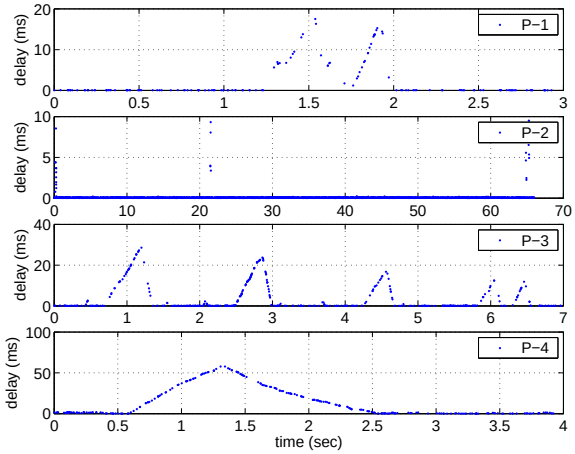


Fig. 15. $\{d^{var}\}$ between link 12 and link d for the periods of P-1 to P-4

magnitude larger than the 99th percentile and affect the shape of the tail, as witnessed by P-1 and P-3 of Figure 13. Such very large delays occur in clustered peaks, caused by short-term traffic bursts that lasts only a few seconds that are not captured by the 5-minute average link utilization as reported by SNMP data.

6. SNMP data exploration

In the previous section we showed that the peaks in the variable delay correlate with utilization levels greater than 90% on the bottleneck link B. This link corresponds to a long-haul link that connects two PoPs inside the Sprint IP backbone network. Given that such high utilization levels are highly unusual in an operational IP backbone network, in this section we look into the reasons leading to such a phenomenon. To address this topic we use SNMP data in conjunction with intra- and inter-domain information.

In Figure 16 we present the throughput measurements collected every 5 minutes for link B throughout the month of November. We notice that link B experiences a jump in its utilization on November 15th, 2002. Traffic on link B increased by 600 Mbps on that particular date and remains at this level until November 21st, 2002 (when our packet trace collection took place). Subsequent days show a significant drop in traffic that stabilizes around 600 Mbps. In what follows we explore the reasons behind the abrupt changes in throughput observed for link B.

Increase in traffic on link B is caused by the multiplexing of the traffic incoming to the router where

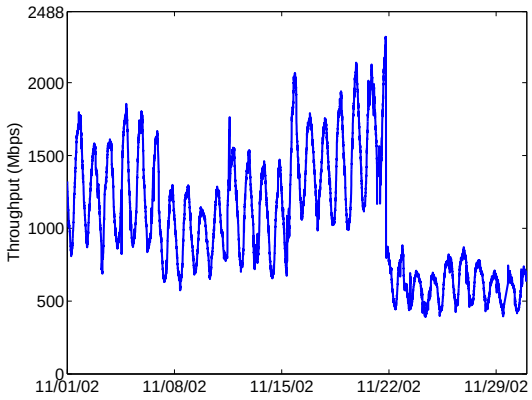


Fig. 16. Throughput measurements for link B throughout the month of November 2002

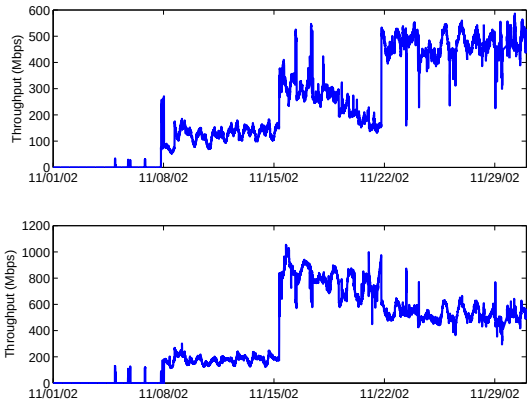


Fig. 17. Throughput measurements for outgoing link of edge router

link B belongs. Looking into the configuration information for that particular router, we find that it serves 26 links. Increase in the utilization of the output link B could be due to two reasons: (i) increase in the overall incoming traffic, or (ii) a jump in the utilization of one of the input links. Given that the increase observed occurs as a jump, the second option may be more likely and lead to the incoming link causing the surge in traffic. To address this issue, we look into the utilization of all the input links for the same period of time and succeed in identifying a single link that experiences a jump of equivalent activity.

Backtracking to the router where this link is attached to, we can repeat the same analysis and attempt to identify one of the input links to this router that manifest the same 600 Mbps jump on November 15th, 2002. It turns out that performing this operation we always manage to identify one single input link manifesting the same behavior. In that way, we arrive at the access router responsible for the identified surge of traffic. This router is located on the west coast of the United States and connects two customers, at OC-12 and Gigabit Ethernet speeds. The first customer is a large Tier-2 provider, and the other one is a large Internet provider in Asia. The SNMP repository used for this analysis does not include information for customer links but looking at the 2 OC-192 links connecting this router to the backbone routers inside the same PoP, we can notice the increase in the total outgoing traffic. In Figure 17 we present the utilization of the two output links of that particular access router.

Indeed, similar jumps in the utilization of both links can be observed for that same date. This increase could be due to one or both customer, but lack of SNMP information for the customer links does not allow for such a task. Nevertheless, we can correlate the time when we witness the abrupt jump in the output utilization (5 minutes after midnight) to announcements made in the inter-domain routing. We have installed route listeners inside the Sprint network and record every message exchanged in the intra- and inter-domain routing. Looking through the BGP logs, we did find that the 2 customers attached to this edge router did advertise 137, and 714 new network prefixes on that location at this same time. New accessible prefixes through this router could lead to additional traffic transiting that access router. Consequently, the increase we witness in the core of the network on link B is due to changes in the inter-domain routing that took place around midnight on November 15th, 2002. We also note, that link B was the only link that was negatively impacted by this increase in traffic; all other links in the path had adequate capacity to absorb the surge in traffic without reaching prohibitive levels of utilization.

The inter-domain routing change took place 6 days before we measured the large variable delays through the network. Only 6 days later did the bottleneck link reach 90% utilization. In addition, it maintained this high levels of utilization only for 30 minutes. We do see that at approximately 8 pm on November 21st, 2002, there is a sudden drop in the utilization of the bottleneck link and its throughput reduces from 2.3 Gbps down to 800 Mbps.

Link B is an inter-PoP link connecting two PoPs on the east coast of the United States. These two PoPs were interconnected through 4 long-haul links. Before November 21st, 2002, only 2 of those links

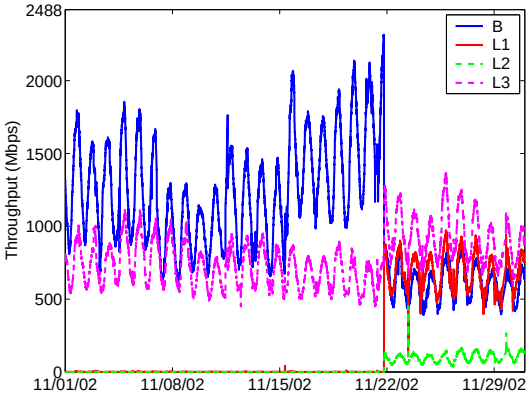


Fig. 18. Throughput of the PoP interconnecting links

carried significant amounts of traffic. After the time period, when we observed congestion, we see that the distribution of traffic across the 4 links has changed. Three of the links carry approximately equal amounts of traffic, while the fourth link ($L2$) now carries approximately 200 Mbps. These results are presented in Figure 18. Such a change in behavior could be due to changes in the ISIS weights for links $L2$ and $L3$. Using the logs of our ISIS router listener for that period of time, we did confirm that such a change in the traffic observed between the two PoPs was due to changes in the intra-domain routing configuration.

In summary, using SNMP and intra- and inter-domain routing information, we showed that: (i) the increase in utilization at the bottleneck link is due to changes in the inter-domain routing, that caused a surge of traffic at an access router on the west coast of the United States, (ii) high levels of utilization only persisted for 30 minutes. Network operations resorted to changes in the ISIS configuration for the better load balancing of the traffic between the two PoPs, that link B interconnected. Consequently, performance degradation occurred for a small period time and was averted immediately.

7. Summary

In this work we present a step-by-step analysis of point-to-point delay from an operational tier-1 backbone network. We first isolate the fixed components in delay, namely, propagation delay, transmission delay, and per-packet overhead at the router. When there are more than one path between two points, we determine the path of each packet, using the minimum delay for each 2-tuple flow and the TTL delta. Once we identify a set of packets that followed the same path, we obtain the minimum path transit time per packet size and subtract it from the point-to-point delay to obtain the variable delay. When the link utilization on all links of the path is under 90%, the 99th percentile of the variable delay remains under 1 ms over 4 to 5 hops and is thus insignificant. However, when a link on the path is utilized over 90%, it becomes a bottleneck in the sense that the weight of the variable delay distribution shifts and even the 90th percentile of the variable delay reaches above 500 μs and the 99th percentile well beyond 1 ms.

Though rare and few in numbers, there are peaks in variable delay often reaching tens of milliseconds in magnitude. We observe such peaks even at below 90% of bottleneck link utilization. We show that these peaks do affect the tail shape of the distribution (above the 99.9th percentile point). We have also investigated the reasons that caused high traffic load in the bottleneck link. With extensive exploration of SNMP data and intra- and inter-domain routing information, we have found that it is due to changes in the inter-domain routing that caused a surge of traffic at an access router.

As we see in this work, many factors contribute to the point-to-point delay in the network. In our observation, point-to-point delays differ as much as 6 ms due to ECMP. Depending on the problem at hand, network engineers can improve current practices by changing a relevant factor.

We believe our observations will be beneficial for network and protocol design in future. Particularly, our work sheds light on designing effective delay monitoring schemes [2]. For instance, when using active probes for monitoring purpose, network operators and managers should make sure that those probes cover all ECMPs, or at least the longest path, so that they stay informed about the worst-case scenario of their network performance.

References

[1] J.-C. Bolot. End-to-end packet delay and loss behavior in the Internet. In *Proceedings of ACM SIGCOMM*, San Francisco, August 1993.

[2] B.-Y. Choi, S. Moon, R. Cruz, Z.-L. Zhang, and C. Diot. Practical delay measurement for ISP. In *Proceedings of ACM CoNEXT*, Roulouse, France, October 2005.

[3] K. C. Claffy, T. Monk, and D. McRobb. Inter-

net tomography. *Nature*, 1999.

[4] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and C. Diot. Packet-level traffic measurements from the Sprint IP backbone. *IEEE Network*, 17(6):6–16, November-December 2003.

[5] U. Hengartner, S. Moon, R. Mortier, and C. Diot. Detection and Analysis of Routing Loops in Packet Traces. In *Proceedings of ACM/SIGCOMM Internet Measurement Workshop*, Marseilles, France, Nov 2002.

[6] G. Huston. *ISP Survival Guide: Strategies for Running a Competitive ISP*. John Wiley & Sons, October 1998.

[7] K. Papagiannaki, S. Moon, C. Fraleigh, P. Thiran, and C. Diot. Measurement and analysis of single-hop delay on an IP backbone network. In *Proceedings of INFOCOM*, San Francisco, April 2002.

[8] V. Paxson. End-to-end routing behavior in the Internet. *IEEE/ACM Transactions on Networking*, 5(5):610–615, November 1997.

[9] V. Paxson. *Measurement and Analysis of End-to-End Internet Dynamics*, chapter 16. Ph.D. Thesis, University of California, Berkeley, April 1997.

[10] R. Ramjee, J. Kurose, D. Towsley, and H. Schulzrinne. Adaptive playout mechanisms for packetized audio applications in wide-area networks. In *Proceedings of INFOCOM*, Montreal, Canada, 1994.

[11] W. Stallings. *SNMP, SNMPv2, SNMPv3, and RMON 1 and 2*. Addison Wesley, 3rd edition, 1999.

[12] D. Thaler and C. Hopps. Multipath issues in unicast and multicast next-hop selection. Internet Engineering Task Force Request for Comments: 2991, November 2000.