Scheme to Measure Packet Processing Time of a Remote Host through Estimation of End-Link Capacity

Khondaker M. Salehin, Roberto Rojas-Cessa, Chuan-bi Lin, Ziqian Dong, and Taweesak Kijkanjanarat

Abstract—As transmission speeds increase faster than processing speeds, the packet processing time (PPT) of a host is becoming more significant in the measurement of different network parameters in which packet processing by the host is involved. The PPT of a host is the time elapsed between the arrival of a packet at the data-link layer and the time the packet is processed at the application layer (RFCs 2679 and 2681). To measure the PPT of a host, stamping the times when these two events occur is needed. However, time stamping at the data-link layer may require placing a specialized packet-capture card and the host under test in the same local network. This makes it complex to measure the PPT of remote end hosts. In this paper, we propose a scheme to measure the PPT of an end host connected over a single- or multiple-hop path and without requiring time stamping at the data-link layer. The proposed scheme is based on measuring the capacity of the link connected to the host under test. The scheme was tested on an experimental testbed and in the Internet, over a U.S. inter-state path and an international path between Taiwan and the U.S. We show that the proposed scheme consistently measures PPT of a host.

Index Terms—Active measurement, intra-probe gap, link capacity, one-way delay, packet processing time, interrupt coalescence.

1 Introduction

Packet processing time (PPT) of a host is the time elapsed between the arrival of a packet in the host's input queue of the network interface card, NIC, (i.e., the data-link layer of the TCP/IP suite) and the time the packet is processed at the application layer [1], [2]. As link rates increase faster than processing speeds [3]–[6], the role of PPT becomes more important in the measurement of different network parameters.

One-way delay (OWD) in a local area network (LAN) is an example of a parameter that PPT can significantly impact [7]. Figure 1 illustrates the OWD of packet P over an end-to-end path, between two end hosts, the source (src) and the destination (dst) hosts. The figure shows the different layers of the TCP/IP protocol stack that P traverses at both end hosts. According to the RFC 2679 [1], the actual OWD is the wire time that the packet experiences in the trip from src to dst. The wire time includes the transmission time (t_t), the queueing delay (t_q), and the propagation time (t_p), or $OWD = t_t + t_q + t_p$, as Figure 1 shows. However, as time stamping of packet creation at src (PPT_{src}) and packet receiving at dst (PPT_{dst}) takes place at the application layer, the coarsely measured OWD may include these PPTs, as an apparent OWD (OWD') of packet P between src and dst, or:

$$OWD' = PPT_{src} + t_t + t_q + t_p + PPT_{dst} \quad (1)$$

Moreover, because of the low transmission rates of legacy systems, PPT has been considered so far negligible (i.e., $PPT_{src} = PPT_{dst} \simeq 0$).

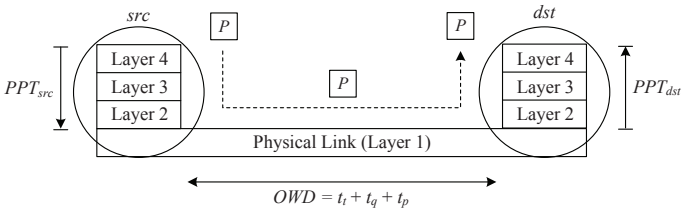


Fig. 1. End-to-end one-way delay (OWD) of packet P over a single-hop path.

As data rates increase, the contribution of PPT on OWD increases, and the error in the measurement of OWD in high-speed LANs can be large if PPTs are neglected. For example, the measurement of OWD between end hosts connected over a 100-Mb/s link using 1500- and 40-byte packets would have errors of 2.5 and 9%, respectively, for $PPT_{src} = PPT_{dst} = 2$ μs, an average $t_q = 40$ μs [8], and $t_p = 0.5$ μs, considering a 100-m Fast-Ethernet cable. In these calculations, error = $|\frac{(OWD - OWD')}{OWD}| \times 100$ %. This error increases to 108% when the queuing delay is relieved (i.e., $t_q \simeq 0$ μs [8]–[13]) for a 40-byte packet, which constitutes 50% or more of the IP traffic [14], [15]. In a similar scenario, this error is 16% on a 1-Gg/s link (as $t_p = 25$ μs for a 5-km optical cable in Gigabit Ethernet [16]). Therefore, PPT must be considered for an accurate measurement of OWD.

Similarly, knowledge of the PPT of servers can be used in financial-trading data centers for identifying which servers

- *K.M. Salehin and R. Rojas-Cessa (Corresponding Author) are with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102, USA. E-mail: {kms29, rojas}@njit.edu*
- *C. Lin is with the Department of Information and Communication Engineering, Chaoyang University of Technology, Wufeng District, Taichung, 41349, Taiwan. Email: cblin@cyut.edu.tw*
- *Z. Dong is with the Department of Electrical and Computer Engineering, New York Institute of Technology, New York, NY 10023, USA. Email: ziqian.dong@nyit.edu*
- *T. Kijkanjanarat is with the Department of Electrical and Computer Engineering, Thammasat University, (Rangsit Campus), Pathumtani, Thailand. Email: taweesak@engr.tu.ac.th.*

comply with the required processing transaction speed. This information would increase customer and service provider confidence in these services [17]–[20].

In a wide area network (WAN), high-resolution OWD measurement can be used to increase accuracy in IP geolocation [21]–[24]. In delay based IP geolocation, each microsecond of propagation delay varies the estimated geographic distance by 200 m between two end hosts connected over optical links [25]. PPT is also an important parameter in the measurement of link capacity and available bandwidth on high speed networks [26], [27]. For example, in schemes for the measurement of link capacity based on packet pairs [28]–[31], 2 μs of PPT in the host involved in the measurement can incur 16% error on a 1-Gb/s link when 1500-byte packets are used in the packet pair. This error increases as the packet length decreases.

The measurement of the PPT of a host can be complex because of the following two reasons: 1) The host must record the time a packet arrives in the data-link layer and the time the application layer processes the packet (here, the time stamping performed at the application layer is considered as the packet-processing event). However, time stamping at the data-link layer is not readily available in typical NICs. Time stamping a packet at the data-link layer for PPT measurement can be performed by placing a specialized packet-capture card in the same subnet where the end host under test is located. 2) The host under test may be remotely located from the source host and access to the host is only allowed through the Internet. These issues raise the following question: is it possible to measure the PPT of remote end hosts through the network?

As a response to this question, we propose an active scheme to measure the PPT of remotely located hosts. The proposed approach consists of the following three components: 1) An active probing scheme to estimate the capacity of the link directly connected to the remotely located host, called the *end link*, using pairs of packets to collect data samples (i.e., gap values in between the pairs of packets), which contain PPT information of the host. 2) A methodology to detect and remove sampled data affected by the network traffic and other network phenomena. 3) A methodology to obtain the PPT of the host under test from the useful collected data. The proposed approach is the first method to remotely measure the PPT of a host, to the best of our knowledge.

The remainder of the paper is organized as follows: In Section 2, we introduce the proposed scheme to measure PPT of a remote end host. In Section 3, we show experimental results of the proposed scheme tested on a testbed and in the Internet. In Section 4, we analyze the data samples obtained in the experiments. In Section 5, we present an analytical model for determining the size of the packets in a compound probe. In Section 6, we discuss existing schemes on link-capacity measurements. In Section 7, we discuss the effect of different parameters of the proposed scheme. In Section 8, we present the conclusions of this work.

2 Scheme for PPT Measurement

The hypothesis of this work is that PPT of an end host can be measured remotely, without physical access to it, through the estimation of the end-link capacity, which is the link directly connected to the host. Here, link capacity is equivalent to transmission speed of the link. The end-link capacity is estimated by using *compound probes* [32]–[34], which are sent from a source host to the remote host, which is the host under test. The compound probe consists of two packets: a heading packet (P_h) and a trailing packet (P_t). Because a single time stamp is issued per packet arrival (at the time the last bit of the packet arrives), two packets are needed to estimate the transmission time of the P_t packet. The transmission time could be underestimated if the packets are not back to back, or if there is a dispersion gap between them. Figure 2 shows a compound probe (a) without a dispersion gap and (b) with a dispersion gap. Here, the dispersion gap is defined by the separation between the last bit of P_h and the first bit of P_t. To measure the transmission time of P_t, the dispersion gap must be zero. The compound probe shares some similarities to the typical packet-pair model, which consists of two packets with equal size, used in several other schemes for the measurement of link capacity and available bandwidth [28], [29], [35], [36]. However, our compound probe uses two packets of different sizes, as in the packet-tailgating technique [30], [37].

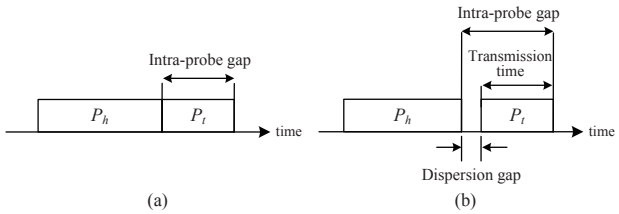


Fig. 2. Compound probe (a) without and (b) with a dispersion gap.

The estimation of end-link capacity (c_n) uses two compound probes with different P_t sizes, $s_t = \{s_1, s_2\}$. The estimated link capacity is used to determine the expected intra-probe gap, $G(s_t)$, as Figure 3 shows. In turn, we use $G(s_t)$ for estimating the offset of the curve, which is the difference of the PPT experienced by P_h and P_t (ΔPPT) at the host. The proposed scheme requires the host to be cooperative by allowing access to the time stamping performed at the application layer.

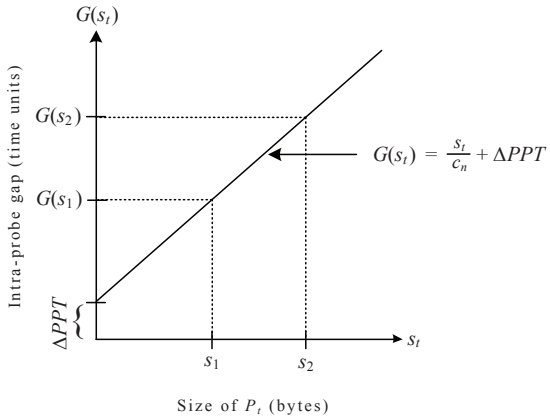


Fig. 3. Linear relationship between the transmission times (intra-probe gap) and trailing packet sizes.

As Figure 3 shows, ΔPPT provides information about

how much delay (which is determined by the CPU and bus speeds, the NIC driver, and the system-call latency [4]) packets experience at the host before being time-stamped. The sizes of the intra-probe gaps of P_ts, $G(s_1)$ and $G(s_2)$, are in (linear) function of the transmission times, and they include ΔPPT. Therefore,

$$G(s_t) = \frac{s_t}{c_n} + \Delta PPT \quad (2)$$

where $\Delta PPT = PPT_t - PPT_h$. ΔPPT is the intersection of the straight line, with slope $\frac{1}{c_n}$, and the y axis; at $s_t = 0$.

For generality, we consider that PPT_h may be equal to or different from PPT_t. Figure 4 shows the case where PPT_h and PPT_t are different. In this figure, TS_h and TS_t are the time stamps for P_h and P_t, respectively, assigned at the application layer.

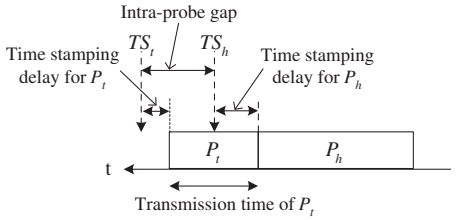


Fig. 4. Variation in time stamping of P_h and P_t by the application layer.

2.1 Measurement Scheme

The proposed methodology to measure PPT is divided into two phases. Phase 1: estimation of the average ΔPPT, ΔPPT_{avg}, from the measured $G(s_t)$s (Figure 6) and Phase 2: estimation of PPT_h and PPT_t using our packet receiving model (Figure 7). Considering the multiple-hop path between src and dst in Figure 5, the detailed steps of the proposed scheme are presented below.

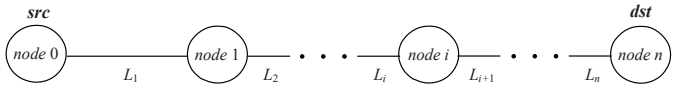


Fig. 5. An n-hop end-to-end path.

Phase 1: Estimation of ΔPPT_{avg}.

1. Send a train of compound probes from src to dst using a P_h with s_h equal to the path's Maximum Transmission Unit (MTU), and a P_t with $s_t = s_1$, where $s_1 < s_h$, such that a large packet-size ratio, $\alpha = \frac{s_h}{s_t}$, in the compound probe is obtained. As an alternative, s_1 can be estimated if information about the capacity of each link along the path is known, or else, it can be determined by exploration of the path (as discussed in Section 5).
2. Send a train of compound probes with $s_t = s_2$, where $s_1 < s_2 < s_h$. The minimum difference between s_1 and s_2 is determined by the resolution of the clock used in the measurement, or in incremental steps of $2Rc_n$ bits (or bytes), where R is the clock resolution (e.g., if $R = 1$ μs and c_n=100 Mb/s, the step is 200 bits).

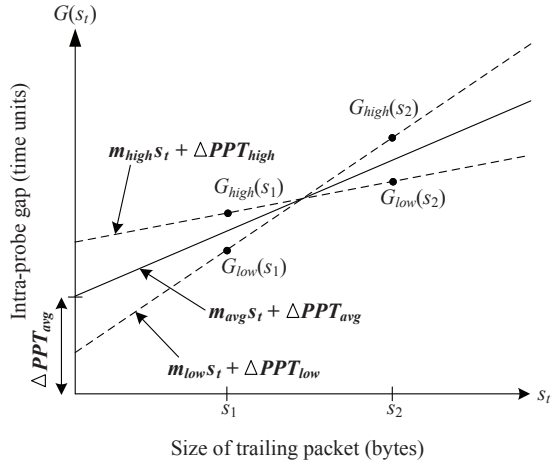


Fig. 6. Estimation of ΔPPT_{avg} using the smallest and the largest intra-probe gaps.

3. Filter out the affected intra-probe gaps and identify the smallest, the largest, and the average intra-probe gaps, $G_{low}(s_t)$, $G_{high}(s_t)$, and $G_{avg}(s_t)$, respectively, for s_1 and s_2.
4. Calculate the smallest and largest values of the reciprocal of the end-link capacity (i.e., slope value) using the measured largest and smallest intra-probe gaps:

$$m_{low} = \frac{G_{high}(s_2) - G_{low}(s_1)}{s_2 - s_1} \quad (3)$$

and

$$m_{high} = \frac{G_{low}(s_2) - G_{high}(s_1)}{s_2 - s_1}. \quad (4)$$

5. Determine the average of the reciprocal of the end-link capacity:

$$m_{avg} = \frac{m_{low} + m_{high}}{2}. \quad (5)$$

6. Calculate the expected intra-probe gap **($\hat{t}_t$)** of a probing train using m_{avg}:

$$\hat{t}_t(s_1) = m_{avg}s_1 + \gamma \quad (6)$$

or

$$\hat{t}_t(s_2) = m_{avg}s_2 + \gamma. \quad (7)$$

where γ is the Interframe Gap (IFG) [38] on the end link, if an Ethernet link is connected to the receiving end host.

7. Estimate ΔPPT_{avg} at dst using the difference between $G_{avg}(s_t)$ and $\hat{t}_t(s_t)$ obtained from Steps 3 and 6, respectively:

$$\Delta PPT_{avg} = G_{avg}(s_1) - \hat{t}_t(s_1) \quad (8)$$

Phase 2: Estimation of PPTs. Figure 7 shows the relationships among the transmission times, intra-probe gaps, and PPTs of P_h and P_t. The receiving buffer at dst uses a first-in-first-out (FIFO) queuing model [35], [36]. The figure shows the timing of P_hs and P_ts of the two compound probes. Because P_hs of the two compound probes have the same length and they arrive before P_ts, the timings of the two P_hs overlap in the figure.

From this model, PPT_ts are determined as:

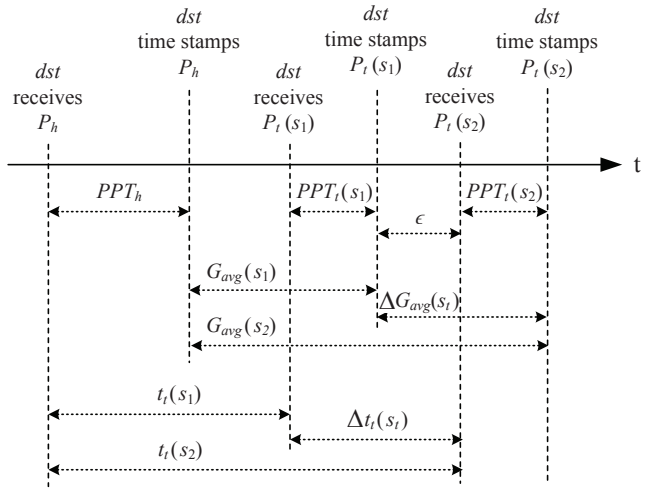


Fig. 7. Packet receiving model at dst.

$$PPT_t(s_1) = PPT_h + G_{avg}(s_1) - t_t(s_1) \tag{9}$$

and

$$PPT_t(s_2) = G_{avg}(s_2) - G_{avg}(s_1) - \epsilon, \tag{10}$$

where ϵ is the interval between the time stamping of $P_t(s_1)$ at the application layer and the arrival of $P_t(s_2)$ at the data-link layer.

On the other hand, if the dispersion gaps of the compound probes are zero and the time-stamping latency at the application layer is smaller than the time to transfer a minimum packet size (e.g., 64 bytes for Ethernet), (5) complies with:

$$m_{avg} = \frac{\Delta G_{avg}(s_t)}{s_2 - s_1} = \frac{\Delta t_t(s_t)}{s_2 - s_1} \tag{11}$$

According to Figure 7, the magnitude of PPT_ts at dst is defined by ϵ, $\Delta t_t(s_t)$, and $\Delta G_{avg}(s_t)$:

$$\Delta t_t(s_t) = PPT_t(s_1) + \epsilon \tag{12}$$

and

$$\Delta G_{avg}(s_t) = PPT_t(s_2) + \epsilon \tag{13}$$

which lead to equal PPTs for both trailing packets, or

$$PPT_t(s_1) = PPT_t(s_2) = PPT_t \tag{14}$$

Therefore, (14) indicates that the magnitude of the two PPT_ts, depends on ϵ, and the largest PPT_t is found when $\epsilon = 0$, or

$$PPT_t = G_{avg}(s_2) - G_{avg}(s_1) = \Delta G_{avg}(s_t) \tag{15}$$

and PPT_h is:

$$PPT_h = PPT_t - \Delta PPT. \tag{16}$$

2.2 Keeping a Zero-Dispersion Gap in a Compound Probe

Gap dispersion (i.e., the increment of the dispersion gap) in a compound probe can occur because of the following two events: 1) one or more cross-traffic packets are inserted between P_h and P_t or 2) if the packet-size ratio (α) is smaller than the link-capacity ratio of a network node i, $cr_i = \frac{c_{i+1}}{c_i}$, where c_i and c_{i+1} are the capacities of the input link, L_i, and output link, L_{i+1}, respectively, of node i. Here, we mainly focus on event 2 (we describe a method to deal with dispersion caused by cross-traffic in Section 2.3). In a network node, if the transmission time of P_h on the output link is smaller than the transmission time of P_t on the input link, the compound probe experiences dispersion [30], [37]. Therefore, the packet-size ratio between P_h and P_t to keep a zero-dispersion gap or to avoid dispersion in node i must follow:

$$\alpha \geq \frac{c_{i+1}}{c_i}. \tag{17}$$

Note that the zero-dispersion gap requirement might be achieved even if the dispersion gap becomes non-zero along the path but it is reduced to zero before reaching the end link (e.g., due to a link-capacity ratio smaller than 1 or queueing at the intermediate nodes). However, we aim at keeping a zero-dispersion along the path, using a suitable α. The condition in (17) is extended for an n-hop path in Section 5.

2.3 Filtering of Affected Gaps

The intra-probe gap of a compound probe can be affected by cross traffic in the measurement path. Affected (i.e., dispersed) intra-probe gaps add errors to the estimation of the end-link capacity. Therefore, we introduce a filtering scheme to detect and remove the affected gaps from the collected samples. In addition to the effect of cross traffic, packet-processing jitter (i.e., the variations of PPT) may also add errors to the measured intra-probe gaps. To remove those errors, the filtering scheme first identifies the smallest and the most frequent intra-probe gaps in a sampled set to determine the level of packet-processing jitter, and it then calculates the standard deviation of the sampled set to find a range of acceptable intra-probe gaps. The use of the most-frequent data element has been considered in link-capacity measurement [32], [37]. The following steps detail the filtering algorithm:

1. Identify the smallest intra-probe gap, $G_{small}(s_t)$, of the sampled set.
2. Determine the frequencies of intra-probe gaps (i.e., histogram) in the sampled set and select the smallest intra-probe gap with the highest frequency $G_{peak}(s_t)$.
3. Estimate the intra-probe gap variations, or packet-processing jitter, as $J = G_{peak}(s_t) - G_{small}(s_t)$, and discard all data elements in the sampled set that are greater than $G_{peak} + J$.
4. Calculate $G_{avg}(s_t)$ and standard deviation σ of the new sampled set.
5. Determine the lower and the upper bounds of the intra-probe gaps as

$$G_{low}(s_t) = G_{avg}(s_t) - \sigma \tag{18}$$

and

$$G_{high}(s_t) = G_{avg}(s_t) + \sigma, \tag{19}$$

respectively.

3 Experimental Results

We measured the PPT of two workstations, a Dell Dimension 3000 workstation (D3000) and a Dell Inspiron I531S workstation (I531S), using the proposed scheme on a controlled

testbed, under different network conditions, and in the (unrestricted) Internet. The goal of the experiments on the testbed is to determine the accuracy of the proposed method in a controlled environment, where link capacities and cross traffic loads are known. The goal of the Internet experiments is to test the proposed scheme under the actual application environment, where link capacities and traffic dynamics are unpredictable and unknown. We compare the outcomes of these two experiments. Table 1 shows the specifications of the workstations.

TABLE 1
End-Host Specifications

	Dell Dimension 3000	Dell Inspiron 531S
Name	D3000	I531S
Processor	Intel Pentium 4	AMD Athlon 64 X2 Dual Core
CPU speed	3 GHz	1 GHz
RAM	512 MB	1024 MB
RAM speed (data width)	400 MHz (64 bits)	667 MHz (64 bits)
PCI bus speed	266 MB/s	133 MB/s
NIC speed	10/100 Mb/s	10/100 Mb/s
Linux kernel version	2.6.18	2.6.18

3.1 Experimental Measurements on a Controlled Testbed

The testbed (Figure 8) was implemented using four different configurations, listed in Table 2. The testbed configurations, each denoted as Cx-r, where x is the index of each of the four considered configurations and r is the capacity of the end link of configuration x, are C1-10, C2-100, C3-100, and C4-100. These configurations provide variations of the end-link capacities, link-capacity ratios, and the location of the narrow link (i.e., the smallest link capacity of the path).

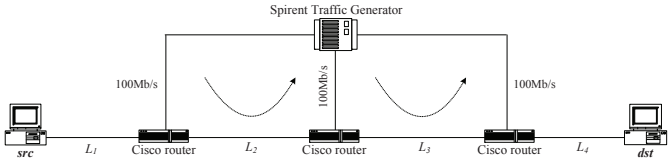


Fig. 8. Testbed setup.

TABLE 2
Testbed-Path Configurations

Path	Link capacity (c_i) (Mb/s)				Link-capacity ratio ($cr_i = \frac{c_i+1}{c_i}$)			Packet-size ratio ($\alpha = \frac{s_h}{s_t}$)	
	c_1	c_2	c_3	c_4	cr_1	cr_2	cr_3	Calculated value	Evaluated value
C1-10	100	155	100	10	1.55	0.645	0.1	1	1
C2-100	10	155	10	100	15.5	0.064	10	10	10
C3-100	100	10	155	100	0.1	15.5	0.645	6.49	6.67
C4-100	10	10	155	100	1	15.5	0.645	6.49	6.67

The testbed consists of one Cisco 3600 router, two Cisco 7200 routers, a Spirent Smartbits 6000C traffic generator, a sender workstation src, and a receiver workstation dst. The D3000 and I531S workstations were used as dst or systems under test. The proposed scheme was implemented as an application on a Linux system, which provides a clock with 1-μs resolution for time stamping (using the pcap library [39]).

We set symmetrical cross-traffic loads between 0 and 90 Mb/s, with steps of 10 Mb/s, on the second and third links of the testbed path (indicated by the dotted-line arrows in Figure 8). No cross-traffic load was applied to the end links. The packet size of each constant-bit-rate (CBR) cross-traffic flow was set between 64 and 128 bytes. The packet sizes are used to generate different levels of traffic loads. We consider that traffic models with different distributions might not differ significantly from CBR traffic at these high loads.

For the compound probes, s_h was determined by the testbed path's MTU, which is 1448 bytes of User Datagram Protocol (UDP) payload plus 54 bytes of encapsulation over the Ethernet links (the Ethernet encapsulation includes a 12-byte preamble, start of frame delimiter, SFD, and frame check sequence, FCS) or a total frame length of 1502 bytes. Here, $s_1 = 87$ bytes and $s_2 = 112$ bytes, which set large packet-size ratios, $\alpha_1 = \frac{1502 \text{ bytes}}{87 \text{ bytes}} = 17.26$ and $\alpha_2 = \frac{1502 \text{ bytes}}{112 \text{ bytes}} = 13.41$, respectively, and they are within the upper bounds of the packet-size ratios found for the evaluated testbed configurations, as shown in $Calculated\ value$ column of Table 2. Section 5.3 discusses further details on the upper bounds of the packet-size ratio. The time stamps of the probe packets at dst were obtained through Wireshark [40]. We tested each configuration and workstation using 500 compound probes and repeated each test 10 times.

Table 3 shows PPT_h, PPT_t, and PPT_{avg} (i.e., $PPT_{avg} = \frac{PPT_h + PPT_t}{2}$) of each workstation measured in the testbed experiments. According to the table, PPT_{avg}s of the D3000 and I531S workstations with 10-Mb/s end link (on C1-10) are 18 and 21 μs, respectively. The PPT_{avg}s of the D3000 workstation with 100-Mb/s end link are 2, 4, and 3 μs on C2-100, C3-100, and C4-100, respectively. Here, the variation in the PPT_{avg} is attributed to the 1-μs clock resolution of the Linux system. For the I531S workstations with 100-Mb/s end link, PPT_{avg} is 3 μs.

TABLE 3
Summary of PPTs of Testbed Measurements

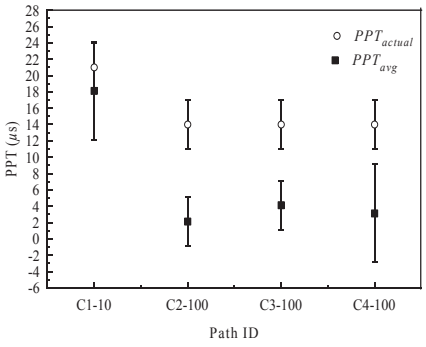
dst	Path	Packet processing time: Actual value PPT_{actual} (μs)	Measured values PPT_h (μs)	PPT_t (μs)	PPT_{avg} (μs)	Error (%)
D3000	C1-10	21	17	19	18	14
D3000	C2-100	14	3	1	2	86
D3000	C3-100	14	6	2	4	71
D3000	C4-100	14	5	1	3	79
I531S	C1-10	16	21	20	21	31
I531S	C2-100	7	3	2	3	57
I531S	C3-100	7	3	2	3	57
I531S	C4-100	7	3	2	3	57

Here, the PPT_{avg}s on the 10-Mb/s end links are larger than those obtained on the 100-Mb/s end links for both workstations. Such large differences are caused by the IFG transmitted in between packets. The time taken by IFGs are also affected by PPT variations. The time difference of IFGs between the two different transmission times is about 9 μs. The link rates then also affect PPT.

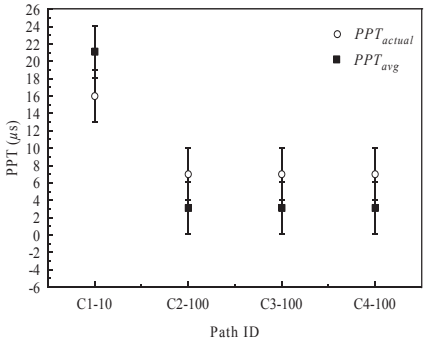
The last column of Table 3 shows the errors of the average PPTs measured on the testbed paths in reference to the actual PPTs (PPT_{actual}) of each workstation, where error = $|\frac{(PPT_{actual} - PPT_{avg})}{PPT_{actual}}| \times 100\%$. PPT_{actual} is measured by a method that places a specialized packet-capture card, Endace DAG 7.5G2 card [41], and the host under test for PPT measurement in the same network segment (i.e., local measurement) [42]. On average, the errors of the PPTs measured

with the proposed scheme are 14% on a 10-Mb/s link and 77% on a 100-Mb/s link for the D3000 workstation, and 31 and 57%, respectively, for the I531S workstation. The results show that the accuracy for 10-Mb/s links is high. The 1-μs of clock resolution of the Linux systems increases the error on the 100-Mb/s links.

Figure 9 shows the measured PPTs of the (a) D3000 and (b) I531S workstations along with their 98-percentile confidence intervals. The hollow circles are PPT_{actual}s and the solid rectangles are the PPT_{avg}s. Figure 9(b) shows that the PPTs of the I531S workstation, measured by the proposed scheme, are very close to the actual values as the overlaps of PPT_{actual}s and PPT_{avg}s show. The measured values for the D3000 workstation, Figure 9(a), are close to the actual values only for 10 Mb/s as this workstation may have a larger variation in its PPT.



(a) D3000



(b) I531S

Fig. 9. Measured PPTs and 98-percentile confidence intervals of the (a) D3000 and (b) I531S workstations on testbed.

We also verified zero-dispersion gaps in the compound probes on the testbed configurations. For this, we used the specialized packet-capture card (Endace DAG card) at dst.

3.2 Experimental Measurements in the Internet

To evaluate the proposed scheme in real network scenario, we performed PPT measurements in the Internet, which consists of end-to-end paths with different link capacities and random traffic loads [29], [43], [44]. In these measurements, we used two Internet paths, a local path in the U.S. and an international path between Taiwan and the U.S., between December 2010 and January 2011. The local path was set from New York Institute of Technology (NYIT), New York, New York, to New Jersey Institute of Technology (NJIT), Newark, New Jersey, and the path is labeled as NYNJ. This path comprises 19 hops. The international path was set between Chaoyang University of Technology (CYUT), Taichung, Taiwan, and NJIT, and it is labeled as TWNJ. This path comprises 21 hops. We configured the workstations at NYIT and CYUT as src nodes, and the nodes at NJIT, the same workstations used in the testbed experiments, as dst nodes. For these experiments, the firewalls at sender and receiver networks were disabled to allow the compound probes to go through. As for the compound probes, we use the same s_ts as in the testbed experiments: 87 and 112 bytes, and s_h of 1512 bytes. As in the testbed experiment, trains of 500 compound probes were also used in each of the 10 measurements. The workstations at both ends were connected to either 10-Mb/s or 100-Mb/s links.

Table 4 shows PPT_h, PPT_t, and PPT_{avg} of the workstations measured in the Internet experiments. PPT_{avg}s of the D3000 and I531S workstations on both the NYNJ and TWNJ paths, with 10-Mb/s end links, are 24 and 21 μs, respectively. In case of 100-Mb/s end links, PPT_{avg}s of the D3000 workstation on the NJNY and TWNJ paths are 3 and 4 μs, respectively. $PPT_{avg}s$ of the I531S workstation on these two Internet paths with 100-Mb/s end links is 3 μs. The errors of the PPTs measured on the Internet paths in reference to the actual PPTs of the workstations are presented in the last column of Table 4. These errors are similar to those obtained on the testbed.

TABLE 4
Summary of PPTs of the Internet Measurements

dst	Path	Packet processing time				Error
		Actual value	Measured values			
		PPT_{actual} (μs)	PPT_h (μs)	PPT_t (μs)	PPT_{avg} (μs)	(%)
D3000	NYNJ-10	21	27	21	24	14
D3000	TWNJ-10	21	27	21	24	14
D3000	NYNJ-100	14	5	1	3	79
D3000	TWNJ-100	14	6	2	4	71
I531S	NYNJ-10	16	21	20	21	31
I531S	TWNJ-10	16	21	20	21	31
I531S	NYNJ-100	7	3	2	3	57
I531S	TWNJ-100	7	3	2	3	57

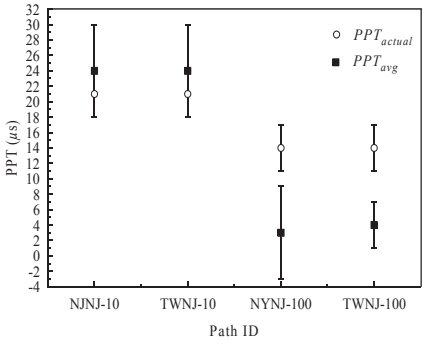
Figure 10 shows the measured PPTs and their 98-percentile confidence intervals of the (a) D3000 and (b) I531S workstations over the Internet paths. This figure shows that the performance of the proposed scheme on both workstations with 10- and 100-Mb/s end links over the Internet paths is consistent with what was measured on the testbed.

4 Quality of the Measured Variables

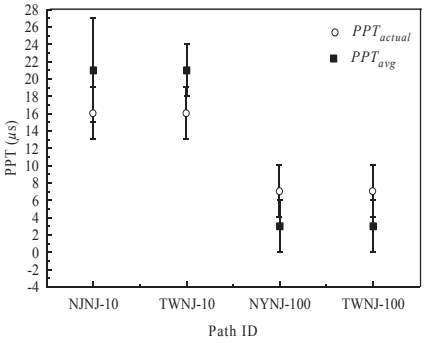
In this section, we present data samples of the measured intra-probe gaps, slope values, and ΔPPTs the testbed and Internet experiments, respectively, to discuss the quality of the estimated end capacities for PPT measurements.

4.1 Data Samples of Testbed Experiments

1) Measured Intra-Probe Gaps without Cross Traffic. Figure 11 shows samples of the distributions of the intra-probe gaps measured on C1-10 and C2-100 at dst without cross-traffic load. The theoretical intra-probe gaps (i.e., $\frac{s_t}{c_4} + \gamma$) for the 87- and 112-byte packets on C1-10 are 80 and 100 μs, respectively, and on C2-100 are 8 and 10 μs, respectively. The



(a) D3000



(b) I531S

Fig. 10. Measured PPTs and 98-percentile confidence intervals of the: (a) D3000 and (b) I531S workstations in the Internet.

theoretical intra-probe gaps on each path are indicated by the solid and dashed vertical-lines in each graph of Figure 11. Each graph also shows the smallest intra-probe gaps with the highest frequency (G_{peak}) for both trailing-packet sizes measured by the workstations. Even though the G_{peak}s are smaller than the theoretical gaps in each graph, the distributions of the measured gaps show that the workstations are not affected by interrupt coalescence [5].

2) Measured Intra-Probe Gaps with Cross Traffic. Figure 12 shows the measured intra-probe gaps (without filtering the affected gaps) on C1-10 and C2-100 with 60% cross-traffic load. Even though Figures 12(b) and 12(d) have some large intra-probe gaps affected by the cross traffic (at the right-hand side of each graph) for both trailing-packet sizes as compared to those of Figures 11(b) and 11(d), the distributions of the intra-probe gaps in Figure 12 are similar to those in Figure 11. The similarity between the measured intra-probe gaps, with and without cross-traffic, suggests that the compound probes are not affected significantly by cross traffic.

Figure 13 shows the distributions of the filtered intra-probe gaps of those in Figure 12. This figure also shows G_{peak}, and the average intra-probe gap, G_{avg}, for both trailing-packet sizes. The distributions of the filtered gaps show that the proposed filtering scheme eliminates the outliers caused by the cross traffic in Figures 12(b) and 12(d).

4) Summary of Intra-Probe Gaps and Quality of End-link Capacity Estimation. Table 5 shows the values of $G(s_t)_{avg}$, m_{avg}, $\hat{t}_t$, and ΔPPT_{avg} measured on the D3000 and I531S workstations in the testbed experiments. As the table shows, the measured ΔPPT_{avg} of the D3000 and I531S workstations on C1-10 are 2 and -1 μs (the negative sign means that $PPT_h > PPT_t$), respectively. The measured ΔPPT_{avg} of these workstations on C2-100, C3-100, and C4-100 are -2, -4, and -4 μs, respectively, for the D3000 workstation, and -1 μs for the I531S workstation.

TABLE 5
Summary of Intra-Probe Gaps of Testbed Experiments

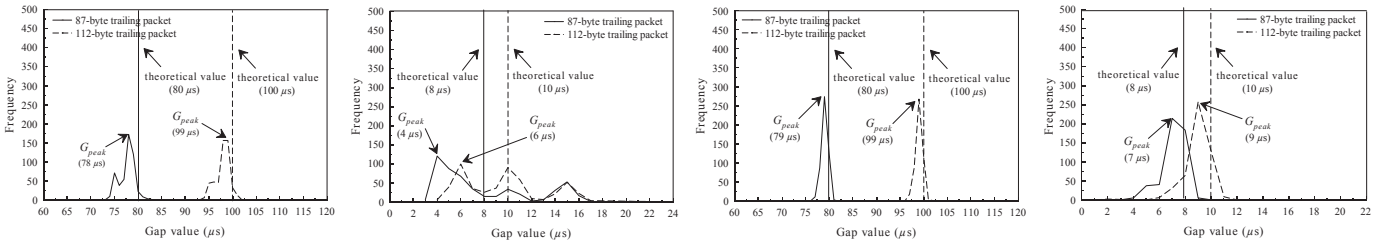
dst	Path	Packet size	Measured intra-probe gap $G(s_t)$			Slope value m		Expected intra-probe gap	Processing time ΔPPT	
		s_t	$[low, high]$	avg	Δ	$[low, high]$	avg	$\hat{t}_t$	avg	std
D3000	C1-10	87	[77, 80]	79	19	[0.92, 0.64]	0.78	77.46	2	1.5
		112	[96, 100]	98				96.96		
D3000	C2-100	87	[4, 5]	5	1	[0.12, 0]	0.06	6.18	-2	0.7
		112	[5, 7]	6				7.68		
D3000	C3-100	87	[3, 5]	4	2	[0.16, 0]	0.08	7.92	-4	0.5
		112	[5, 7]	6				9.92		
D3000	C4-100	87	[4, 6]	5	1	[0.12, -0.04]	0.08	7.92	-4	1.7
		112	[5, 7]	6				9.92		
I531S	C1-10	87	[78, 80]	79	20	[0.88, 0.72]	0.8	79.2	-1	0.9
		112	[98, 100]	99				99.2		
I531S	C2-100	87	[6, 8]	7	2	[0.16, 0]	0.08	7.92	-1	1.0
		112	[8, 10]	9				9.92		
I531S	C3-100	87	[6, 8]	7	2	[0.16, 0]	0.08	7.92	-1	1.0
		112	[8, 10]	9				9.92		
I531S	C4-100	87	[6, 8]	7	2	[0.16, 0]	0.08	7.92	-1	1.0
		112	[8, 10]	9				9.92		

About the quality of estimated end-link capacities, the average slopes (m_{avg}) measured by the D3000 and I531S workstations on C1-10 are 0.78 and 0.8, respectively, and the actual slope (i.e., the expected slope) is 0.8 for c_4 = 10 Mb/s. In the cases of C2-100, C4-100, and C4-100, the actual slope is 0.08 (for c_4 = 100 Mb/s) and the values measured by the D3000 workstation are 0.06 on C2-100, and 0.08 on both C3-100 and C4-100. The slopes measured by the I531S workstation are 0.8 and 0.08 on C1-10 and C2-100 to C4-100, respectively. Table 5 shows that the proposed scheme measures the end-link capacity of each path with high accuracy for PPT measurement, which supports the hypothesis stated in Section 2.

4.2 Data Samples of Internet Experiments

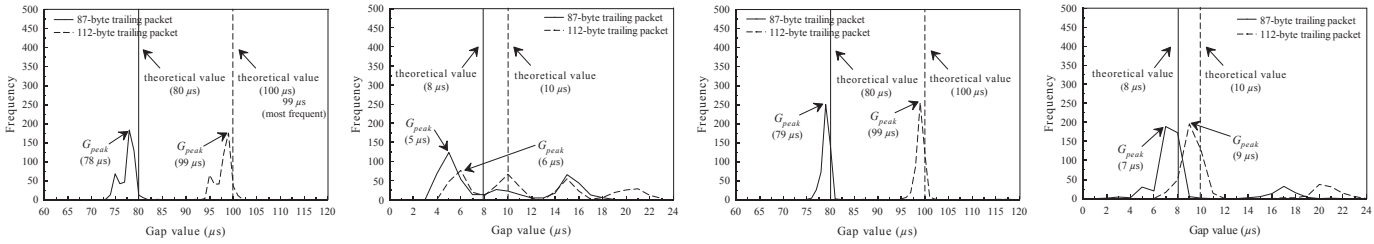
1) Measured Intra-Probe Gaps. Figure 14 shows samples of the intra-probe gaps, measured by the dst nodes with 10-Mb/s and 100-Mb/s end links. The graphs in Figure 14(a)-(d) show the distributions of intra-probe gaps on the NYNJ path, and Figures 14(e)-(h) show the intra-probe gap distributions on the TWNJ path. The distributions of the intra-probe gaps measured by both workstations on the Internet paths are similar to those measured on the testbed. This similarity proves that the compound probe and the used packet sizes are robust against the random traffic and link capacities of Internet paths.

2) Summary of Intra-Probe Gaps and Quality of End-link Capacity Estimation. Table 6 shows the values of $G(s_t)_{avg}$, m_{avg}, $\hat{t}_t$, and ΔPPT_{avg} measured on the D3000 and I531S workstations in the Internet experiments. ΔPPT_{avg} of the D3000 workstation on the 10- and 100-Mb/s end links are -6 and -4 μs, respectively. ΔPPT_{avg} of the I531S workstation on both the 10- and 100-Mb/s end links are -1 μs. As for the estimated end-link capacities, m_{avg}s measured by the D3000 workstation on the Internet paths are 0.84 and 0.08 when the end-link capacities are 10- and 100-Mb/s, respectively. These values are measured as 0.8 and 0.08 on the respective end-link capacities by the I531S workstation. These values show that the proposed scheme consistently estimates end-link capacity in the Internet with a high accuracy.



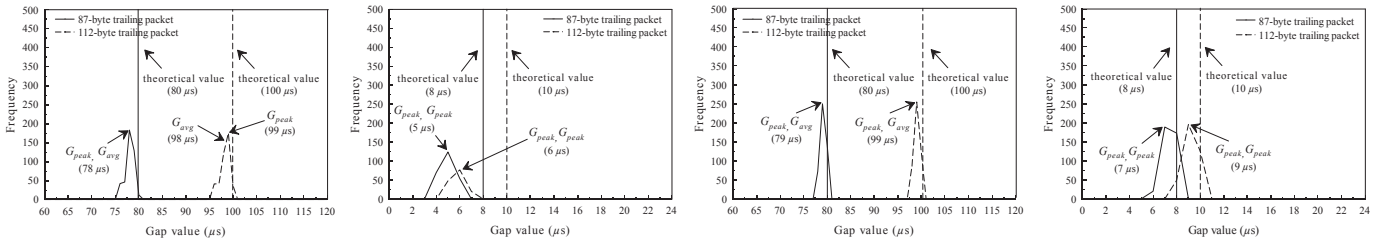
(a) D3000 on C1-10 (b) D3000 on C2-100 (c) I531S on C1-10 (d) I531S on C2-100

Fig. 11. Distributions of intra-probe gaps, with no cross-traffic load in the network, measured by the D3000 workstation on: (a) C1-10 and (b) C2-100, and by the I531S workstation on: (c) C1-10 and (d) C2-100.



(a) D3000 on C1-10 (b) D3000 on C2-100 (c) I531S on C1-10 (d) I531S on C2-100

Fig. 12. Distributions of intra-probe gaps, under 60% cross-traffic load, measured by the D3000 workstation on: (a) C1-10 and (b) C2-100, and by the I531S workstation on: (c) C1-10 and (d) C2-100.



(a) D3000 on C1-10 (b) D3000 on C2-100 (c) I531S on C1-10 (d) I531S on C2-100

Fig. 13. Distributions of filtered intra-probe gaps, under 60% cross-traffic load, measured by the D3000 workstation on: (a) C1-10 and (b) C2-100, and by the I531S workstation on: (c) C1-10 and (d) C2-100.

TABLE 6
Summary of Intra-Probe Gaps of the Internet Experiments

dst	Path	Packet size s_t	Measured intra-probe gap $G(s_t)$ [low, $high$]	avg	Δ	Slope value m [low, $high$]	avg	Expected intra-probe gap $\hat{t}_t$	Processing time ΔPPT avg	std
D3000	NYNJ-10	87	[75, 79]	77	21	[1, 0.68]	0.84	82.68	-6	1.6
		112	[96, 100]	98				103.68		
D3000	TWNJ-10	87	[75, 79]	77	21	[1, 0.68]	0.84	82.68	-6	1.6
		112	[96, 100]	98				103.68		
D3000	NYNJ-100	87	[4, 6]	5	1	[0.12, -0.04]	0.08	7.92	-4	1.5
		112	[5, 7]	6				9.92		
D3000	TWNJ-100	87	[3, 5]	4	2	[0.16, 0]	0.08	7.92	-4	0.7
		112	[5, 7]	6				9.92		
I531S	NYNJ-10	87	[78, 80]	79	20	[0.72, 0.88]	0.8	79.2	-1	1.3
		112	[98, 100]	99				99.2		
I531S	TWNJ-10	87	[78, 80]	79	20	[0.72, 0.88]	0.8	79.2	-1	0.8
		112	[98, 100]	99				99.2		
I531S	NYNJ-100	87	[6, 8]	7	2	[0, 0.16]	0.08	7.92	-1	0.9
		112	[8, 10]	9				9.92		
I531S	TWNJ-100	87	[6, 8]	7	2	[0, 0.16]	0.08	7.92	-1	0.8
		112	[8, 10]	9				9.92		

5 Packet-Size Ratio for Keeping Zero-Dispersion Gap

The measurement of PPT of a remote end host requires accurate measurement of end-link capacity of an end host. This accuracy depends on having a zero dispersion gap in the compound probe, as discussed in Section 2.2. Zero dispersion in a compound probe can be achieved by choosing suitable P_h and P_t and those are determined by the path link capacities and traffic on the path. Based on [34], we present an analytical model for determining the packet sizes in the compound probe to achieve a zero-dispersion gap considering the above stated phenomenon of a path.

5.1 Size of Probing Packets without Cross-Traffic Effect

Consider that the capacities of the links $L_1, L_2, \ ..., L_n$ between src and dst along the n-hop path in Figure 5 are $c_1, c_2, \ ..., \ c_n$. In this end-to-end path, the end link (L_n) is the narrow link if $c_n \leq c_i$, where $n \geq 2$; otherwise, L_i, where $n \neq i$, constitutes the narrow link of the path.

When the narrow link is located on L_n, the size of P_t is calculated using the following relationship:

$$\frac{s_h}{c_n} - \frac{s_h}{\alpha c_{n-1}} = 0 \tag{20}$$

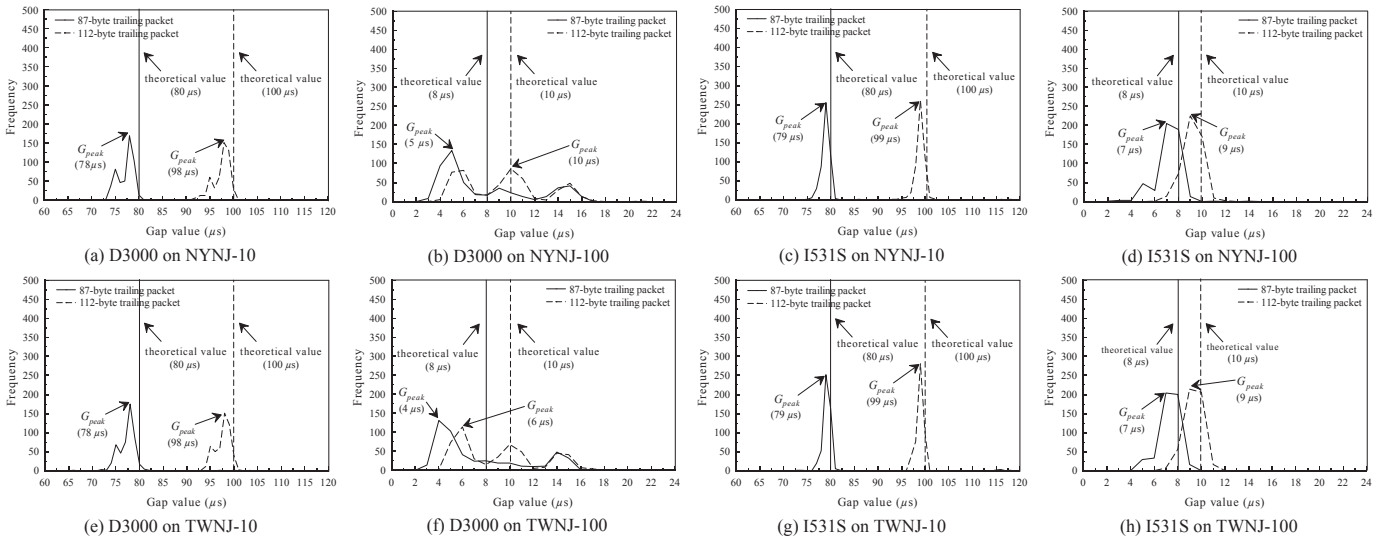


Fig. 14. Distributions of intra-probe gaps measured at dst (NJIT) on NYNJ path: by the D3000 workstation with a (a) 10-Mb/s end link and (b) 100-Mb/s end link, and by the I531S workstation with a (c) 10-Mb/s end link and (d) 100-Mb/s end link; and on TWNJ path: by the D3000 workstation with a (e) 10-Mb/s end link and (f) 100-Mb/s end link, and by the I531S workstation with a (g) 10-Mb/s end link and (h) 100-Mb/s end link.

where $\frac{s_h}{\alpha} = s_t$. The above relationship defines the required condition to ensure a zero-dispersion gap between P_h and P_t at node n. Using (20) and α, the largest size of P_t is:

$$s_t = \frac{s_h}{c_n} c_{n-1} \tag{21}$$

When the narrow link is located on L_i, the relationship between the packet sizes and link capacities required to ensure a zero-dispersion gap between P_h and P_t is:

$$(\frac{s_h}{c_n} - \frac{s_h}{\alpha c_{n-1}}) + (\frac{s_h}{c_{n-1}} - \frac{s_h}{\alpha c_{n-2}}) + ... + (\frac{s_h}{c_{z+1}} - \frac{s_h}{\alpha c_z}) = 0 \tag{22}$$

where c_z is the capacity of a link connected to a node z, such that its link-capacity ratio, $lr_z = \frac{l_{z+1}}{l_z}$, is the largest along the path, located after the narrow link (in the direction from src to dst) and that l_z also is the link closest to dst (e.g., if there are two nodes following the narrow link closest to dst of the path have the largest link-capacity ratio, the node located the closest to dst is selected). Therefore, the index z is such that $1 \leq z \leq (n-1)$. The largest size of P_t in (22) is:

$$s_t = s_h \frac{\sum_{j=z+1}^{n} \frac{1}{c_j}}{\sum_{j=z}^{n-1} \frac{1}{c_j}} \tag{23}$$

5.2 Packet Sizing to Reduce Interference by Cross Traffic

Figure 15 shows an example of a compound probe forwarded from input link L_i to output link L_{i+1} by node i when the compound probe is affected by two cross-traffic packets, denoted by D_h and D_t. The capacity of L_i and L_{i+1} are c_i and c_{i+1}, respectively, where $c_i < c_{i+1}$ in this example. Here, P_h and P_t arrive at node i with a zero-dispersion gap. However, the compound probe experiences dispersion on L_{i+i}, as shown by the dispersion gap in the figure.

The intra-probe gap of a compound probe at the output link of node i is defined as:

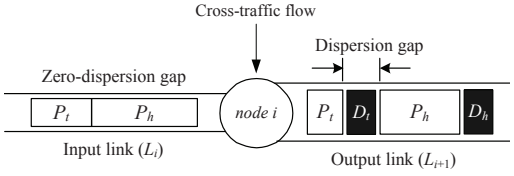


Fig. 15. Forwarding of a compound probe from the input link to the output link by node i in presence of cross traffic.

$$G(s_t)_{i+1} = \frac{s_t}{c_{i+1}} + \delta_{i+1}, \tag{24}$$

where the dispersion gap at the output link, δ_{i+1}, is:

$$\delta_{i+1} = \begin{cases} \delta_i - \Delta tr(h,t)_i & \text{if } \delta_i - \Delta tr(h,t)_i > Qt_i \\ Qt_i & \text{else} \end{cases} \tag{25}$$

In the above equation, δ_i is the dispersion gap at the input link of node i, $\Delta tr(h,t)_i$ is the difference between the transmission time of P_h plus the queuing delay (Qh_i) caused by the cross-traffic packet(s) backlogged ahead of P_h at the output link and the transmission time of P_t at the input link of node i. Qt_i is the increment of the dispersion gap caused by the cross-traffic packet(s) inserted between P_h and P_t at node i. These terms are estimated as:

$$\Delta tr(h,t)_i = \frac{s_h}{l_{i+1}} + Qh_i - \frac{s_t}{l_i} \tag{26}$$

$$Qh_i = \sum_u \frac{\zeta_{u(i)}}{l_{i+1}};\ u \geq 0 \tag{27}$$

$$Qt_i = \sum_v \frac{\zeta_{v(i)}}{l_{i+1}};\ v \geq 0 \tag{28}$$

Here, $\zeta_{u(i)}$ and $\zeta_{v(i)}$ denote the sizes of the cross-traffic packets u and v ahead of P_h and P_t for Qh_i and Qt_i, respectively, at node i.

The measured intra-probe gap by the application layer of node n includes ΔPPT, and it follows (2) and (24):

$$G(s_t)_n = \frac{s_t}{c_n} + \delta_n + \Delta PPT \tag{29}$$

A dispersion in the compound probe due to cross traffic can be detected and the affected gap is discarded by the data-filtering scheme introduced in Section 2.

5.3 Numerical Evaluations of Packet-Size Ratio

We calculated the upper bounds (i.e., the largest possible value) of the packet-size ratios, α, needed to keep the dispersion gap at zero, using (20) and (22), for accurate end-link capacity measurement on the testbed paths, as shown in the ninth column of Table 2. To verify the calculated values, we estimated the end-link capacity (c_4) of each testbed path using different αs based on the dispersion-gap model presented in Section 5.2, considering Qt_i and Qh_i equal to zero in (25) and (26). We estimated c_4 by determining the dispersion gap of a single compound probe on each path where the initial value of s_t was set to 50 bytes and it was gradually increased by 25 bytes until the length of s_h = 1500 bytes is reached to evaluate with different αs. According to Figures 16(a)-16(d), the upper bounds of α, as indicated by the arrow in each graph, are very close to the calculated ones in Table 2. The calculated αs on C1-10 to C4-100 are 1 (s_t =1500 bytes), 10 (s_t =150 bytes), 6.49 (s_t =232 bytes), and 6.49 (s_t =232 bytes), respectively. The evaluated αs in the above stated figures are 1 (s_t = 1500 bytes), 10 (s_t = 150 bytes), 6.67 (s_t = 225 bytes), and 6.67 (s_t = 225 bytes) on C1-10, C2-100, C3-100, and C4-100, respectively. The last column of Table 2 shows these values. Here, the small over estimation of α on C3-100 and C4-100 are produced by the 25-byte step increase of s_t used in the evaluation.

The interference of cross traffic on the measurement of end-link capacity was evaluated considering a uniform cross-traffic load on the second and third links of each path configuration, as indicated by the arrows in Figure 8. In this evaluation, we considered the same αs, as in the no cross-traffic case, along with three different values for u and v, i.e., 1, 5, and 10 packets, in (27) and (28). Figures 16(e)-16(h) show the estimated values of c_4 on C1-10 to C4-100, respectively. The estimated c_4s show that a larger α is required for link-capacity measurement in the presence of cross traffic, except for C1-10, where the end link is the narrow link of the path. For example, when a single cross-traffic packet intrudes the compound probe (i.e., when $u = 1$ and $v = 1$), the lower bound of α increases from 10 to 20 (and s_t decreases from 150 bytes to 75 bytes) on C2-100, and from 6.67 to 8.57 (and s_t decreases from 225 bytes to 175 bytes) on both C3-100 and C4-100, as shown in Figures 16(f)-16(h). Here, we can see that if the narrow link is other than the end link, cross traffic produces dispersion. In such scenario, the largest packet-size ratio achievable on Ethernet is sufficient to ensure a zero-dispersion gap in the considered testbed configurations.

6 Related Work on Measurement of Link Capacity

To the best of our knowledge there is no scheme available for remotely measuring PPT of a host. However, the processing delay of routers have been of interest [45], [8], [46].

Because our proposed scheme is the only method available for remotely measuring the PPT of an end host and the scheme determines PPT from end-link capacity estimation, we limit the discussion to the schemes for measuring link capacity that may be applicable. A plethora of active schemes to measure link capacity has been proposed [10], [28]–[31], [37], [47]–[52]. These schemes can be coarsely classified into two categories: hop-by-hop [10], [30], [50]–[52] and specific-link [37] measurement.

Pathchar, an early hop-by-hop scheme to measure link capacity, is based on measuring the round-trip time, RTT (the traveling time of a packet sent from a source node to the destination node plus the traveling time on the opposite direction) [50]. In this scheme, the network node (i.e., router) directly connected to the source node is used as the first destination, and the capacity of the first link is obtained. The same procedure is applied to the nodes farther away from the source until the remote end host is reached. Pathchar have been reported to produce large errors [4], [10]. There are other schemes that follow similar approach to Pathchar, with, however, different statistical analysis, aimed at reducing probing load [10], [51].

A subsequent hop-by-hop scheme, called Nettimer, uses an approach based on the accumulated delay of an end-to-end path [30]. Nettimer uses a tailgating technique, where two probing packets are sent back-to-back, with a small packet tailgating a large packet, to identify the contribution to the accumulated delay from each single link in the path. The large packet is dropped right before reaching the link of interest and the small packet is left to continue towards the remote end host. The scheme is based on measuring the delays of the small packet. However, the measurement of the path delay using the tailgating technique is affected by large errors in the estimation of link capacity due to cross traffic [30].

Another hop-by-hop scheme, which is resilient to cross traffic and that produces a small probing load, was recently proposed [52]. The scheme is also based on the tailgating technique but each probing packet consists of two small ICMP (Internet Control Message Protocol) packets behind a large data packet, and all sent back-to-back. The time stamps of the ICMP packets, provided by the destination node, are used to measure the intra-probe gap in each probing packet, which are used to estimate the link capacity. Even though the measurement accuracy of the scheme is high, the low resolution of ICMP time stamps (i.e., 1 ms) bounds its applicability to slow link rates.

Different from hop-by-hop measurement, a scheme that measures the link capacity of a target link in a path was proposed [37]. This scheme sends a train of probing packets in pairs, where each pair consists of a large packet tailgated by a small packet, similar to a single pair in Nettimer. All packets in the probing train are dropped before reaching the link of interest while the small packets of the first and last packet pairs reach the remote end host. This scheme measures the gap between the two small packets at the end host of a one-way transmission. The reported accuracy is

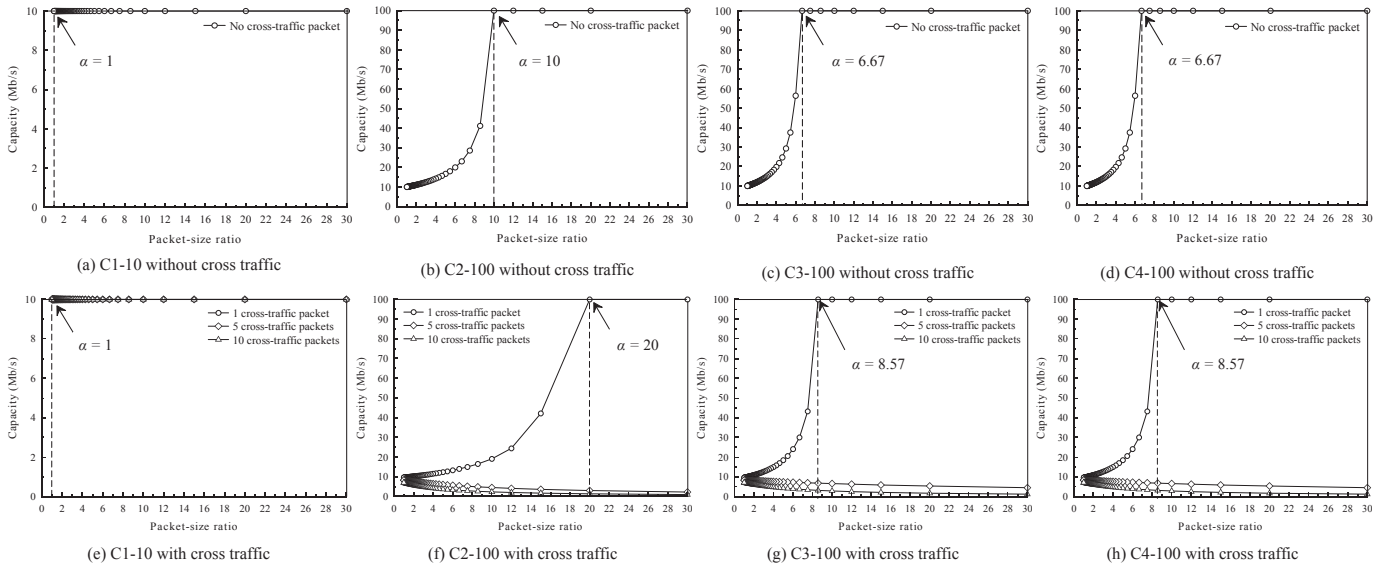


Fig. 16. Evaluated end-link capacities of four different path configurations using different packet-size ratios: (a) C1-10, (b) C2-100, (c) C3-100, and (d) C4-100 without cross traffic. (e) C1-10, (f) C2-100, (g) C3-100, and (h) C4-100 with cross traffic.

higher than that of Nettimer when the measurement path is lightly congested with cross traffic.

7 Discussion

7.1 Error Caused by Clock Resolution

The main limitation of the workstations for the PPT measurement is the low resolution of the clock, as provided by the operating systems. Higher accuracy in the measurement of PPT can be achieved by using a nanosecond-resolution clock, but the implementation of this resolution into the operating systems may degrade the performance of the currently available workstations [4], [5]. Although the two workstations with 100-Mb/s links were tested successfully in the reported experiments, a higher (host) clock resolution than the one used in our experiments is needed to measure PPTs for this link capacity and higher rates. Meanwhile, the proposed scheme can provide a moderate idea of the value of PPT in those cases.

Table 7 shows the largest possible errors ($Error$ column) on the measurement of PPT by 1-μs clock resolution. These errors are not caused by the proposed scheme. As the table shows, the error is up to 57 and 86% for 10 and 100 Mb/s end links. In comparison with the reported values, we see that these errors are equal to or larger than those errors measured in our experimental evaluations except for the I531S workstation with 100-Mb/s end link. Methods to overcome this system drawback can be investigated in the future.

TABLE 7
Maximum Error on PPT due to 1-μs Clock Resolution

dst	Link capacity c_n (Mb/s)	Actual PPT PPT_{actual} (μs)	Measured PPT PPT_{avg} (μs)	Error %
D3000	10	21	9	57
I531S	10	16	9	44
D3000	100	14	2	86
I531S	100	7	4	43

7.2 Internet Link-Capacity Ratios

The Internet experiments were performed without the knowledge of capacities of the intermediate links. Therefore, we recurred to using the two largest packet-size ratios that Ethernet allows rather than exploring the largest link-capacity ratio of the Internet paths. The used packet-size ratios (i.e., $\alpha_1 = 17.37$ and $\alpha_2 = 13.5$) in the compound probes and the consistent measurement results show that the Internet may not have link-capacity ratio(s) larger than the used packet-size ratios. Moreover, the Internet backbone is reported to have link capacities between 1- and 10-Gb/s speeds [27], [53]. Therefore, it is very unlikely to have large link-capacity ratio(s) in an Internet path that could produce dispersion in the intra-probe gap. Even with very large link-capacity ratio(s) along a path (or when approaching to the network core on aggregation links), it is possible to keep a zero-dispersion gap as end links are expected to have small capacities.

7.3 Improved Experimental Measurement of OWD with Application of PPT

As an application of the knowledge of PPT, we performed measurements of OWD in NJIT's production network to show the improved accuracy. We tested OWDs on two end-to-end paths, consisting of 2 and 5 hops. The end hosts were connected to 100-Mb/s end links, except for the 5-hop path where the I531S workstation was connected to a 10-Mb/s end link. The average OWD's measured over the 2- and 5-hop paths are 56 and 216 μs, respectively. Considering the PPTs (i.e., PPT_{actual}s) of the D3000 and I531S workstations, the actual OWDs (i.e., $OWD = OWD' - (PPT_{src} + PPT_{dst})$) over the 2- and 5-hop paths are 35 and 186 μs. These experiments show that knowledge of the PPTs of the workstations reduces the errors by 60 and 16% for the 2- and 5-hop paths, respectively.

7.4 Probing Load

The estimated probing loads used for measuring PPT for the testbed and Internet experiments were 12.89 Mb for 10- and 100-Mb/s links, which was generated by sending 500 pairs of compound probes with an inter-probe space of 100 ms. Therefore, the probing traffic was injected at the rates of 127.52 and 129 Kb/s for 10- and 100-Mb/s links, respectively, which are equivalent to less than 2 and 0.2%, respectively, of the link capacities.

We also measured the effect of probing on the performance of the workstations under test. In this experiment, the compound probes were sent at two inter-probe spaces (i.e., at two different probing rates): 1) 10 ms and 2) 100 ms. The monitored CPU loads on the D3000 and I531S workstations under these two rates were up to 0.5% during the measurement time. Therefore, the loads the proposed scheme generates on the network and on the CPUs of the workstation are small and negligible.

The proposed scheme is able to measure PPT with a smaller number of compound probes. For example, we tested the accuracy the proposed scheme with different numbers of probes and found that the scheme achieves equivalent accuracy by using only 30 or 50 compound probes. The amount of probing traffic and testing time could be small.

8 Conclusions

We proposed a scheme to remotely measure the PPT of an end host (i.e., workstation), which could be connected over a single- or multiple-hop path. To do this, we send two sets of compound probes with different trailing packet sizes and zero-dispersion gaps, from a source host to the remote end host. The intra-probe gaps of the probing packets are used to estimate the end-link capacity. PPT of the remote end host is then determined from the observed deviations of expected end-link capacity measurement.

In the proposed scheme, the zero-dispersion gaps, provided at the generation of the compound probes from the source host, are also used to detect the probes affected by cross traffic along the traveling path. This feature permits filtering the affected gaps and provides immunity against cross traffic. To use this property, we also introduced a model to calculate the dispersion gap under cross traffic, and a filtering scheme to remove intra-probe gaps affected by cross traffic.

We implemented the proposed scheme as a Linux application and it was tested on a controlled testbed and in the Internet for measuring the PPT of two workstations with different specifications. The Internet experiments included two paths: one between New York and New Jersey, and the other between Taiwan and the U.S. (New Jersey). We used the same workstations in both experiment setups and obtained consistent PPTs.

The experimental results show that the proposed scheme achieves high accuracy on 10-Mb/s end links. The accuracy on 100-Mb/s end links decreases as higher clock resolution is desirable; however the PPTs are still measurable using the proposed scheme. Higher clock resolutions may be needed to apply this method on higher link rates. We leave this as future research.

The proposed scheme also proved to be practical for use in the Internet, as the link-capacity ratios of the Internet appear to be no larger than the packet-size ratio that an Ethernet MTU can provide. The consistent accuracy of the proposed scheme on both testbed and Internet paths proves the applicability of the proposed measurement scheme.

References

[1] G. Almes, S. Kalidindi, and M. Zekauskas. RFC 2679 - A one-way delay metric for IPPM. [Online]. Available: http://www.ietf.org/rfc/rfc2679.txt.

[2] ——. RFC 2681 - A round-trip delay metric for IPPM. [Online]. Available: http://www.ietf.org/rfc/rfc2681.txt.

[3] N. McKeown. High performance routers – Talk at IEE, London UK. October 18th, 2001. [Online]. Available: http://tinytera.stanford.edu/ nickm/talks/index.html.

[4] G. Jin and B. Tierney, "System capability effects on algorithms for network bandwidth measurements," in *Proc. of IMC*, FL, USA, 2003, pp. 27–38.

[5] R. Prasad, M. Jain, and C. Dovrolis, "Effects of interrupt coalescence on network measurements," in *Proc. of PAM*, France, 2004, pp. 247–256.

[6] S. Savage. IP router design. [Online]. Available: http://cseweb.ucsd.edu/classes/wi05/cse123a/Lec8.pdf.

[7] A. Hernandez and E. Magana, "One-way delay measurement and characterization," in *Proc. of IEEE ICNS*, Athens, Greece, 2007, p. 114.

[8] K. Papagiannaki, S. Moon, C. Fraleigh, P. Thiran, and C. Diot, "Measurement and analysis of single-hop delay on an IP backbone network," *IEEE Journal of Selected Areas of Communications*, vol. 21, no. 6, pp. 908–921, 2003.

[9] M. Garetto and D. Towsley, "Modeling, simulation and measurements of queuing delay under long-tail Internet traffic," in *Proc. of ACM SIGCOMM*, CA, USA, 2003, pp. 1–11.

[10] A. Downey, "Using pathchar to estimate Internet link characteristics," in *Proc. of ACM SIGCOMM*, MA, USA, 1999, pp. 241–250.

[11] S. Zander and S. J. Murdoch, "An improved clock-skew measurement technique for revealing hidden services," in *Proc. of 17th USENIX Security Symposium*, CA, USA, 2008, pp. 211–225.

[12] V. Paxson, "On calibrating measurement of packet transit times," in *Proc. of ACM SIGCOMM*, WI, USA, 1998, pp. 11–21.

[13] K. Anagnostakis, M. Greenwald, and R. Ryger, "cing: Measuring network-internal delays using only existing infrastructure," in *Proc. of IEEE INFCOM*, CA, USA, 2003, pp. 2112–21 213.

[14] Caida. Packet size distribution comparison between Internet links in 1998 and 2008. [Online]. Available: http://www.caida.org/research/traffic-analysis/pkt_size_distribution/graphs.xml.

[15] R. Sinha, C. Papadopoulos, and J. Heidemann, "Internet packet size distributions: Some observations," USC/Information Sciences Institute, Tech. Rep. ISI-TR-2007-643, May 2007. [Online]. Available: http://www.isi.edu/ johnh/PAPERS/Sinha07a.html

[16] B. Forouzan, *TCP/IP Protocol Suit, ch. 3*. New York, NY, USA: McGraw Hill, 2010.

[17] R. Kompella, K. Levchenko, A. Snoeren, and G. Varghese, "Every microsecond counts: Tracking fine-grain latencies with a lossy difference aggregator," in *Proc. of ACM SIGCOMM*, Barcelona, Spain, 2009, pp. 255–266.

[18] M. Lee, N. Duffield, and R. Kompella, "Not all microseconds are equal: Fine-grained per-flow measurements with reference latency interpolation," in *Proc. of ACM SIGCOMM*, Dehli, India, 2010, pp. 27–38.

[19] J. Sanjuas-Cuxart, P. Barlet-Ros, N. Duffield, and R. Kompella, "Sketching the delay: Tracking temporally uncorrelated flow-level latencies," in *Proc. of IMC*, Berlin, Germany, 2011, pp. 483–496.

[20] Wall street's quest to process data at the speed of light. [Online]. Available: http://www.informationweek.com/news/199200297?pgno=1.

[21] V. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," in *Proc. of ACM SIGCOMM*, CA, USA, 2001, pp. 173–185.

[22] E. Katz-Bassett, J. John, A. Krishnamurthy, T. Anderson, and Y. Chawathe, "Towards IP geolocation using delay and topology measurements," in *Proc. of IMC*, NY, USA, 2006, pp. 71–84.

[23] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of Internet hosts," *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1219–1232, 2006.

[24] Z. Dong, R. Perera, R. Chandramouli, and K. Subbalaksmi, "Network measurement based modeling and optimization for IP geolocation," *Computer Networks*, vol. 56, no. 1, pp. 85–98, 2012.
[25] C. Bovy, H. Mertodimedjo, G. Hooghiemstra, H. Uijterwaal, and P. van Mieghem, "Analysis of end-to-end delay measurements in Internet," in *Proc. of PAM*, CO, USA, 2002, pp. 1–8.
[26] Internet2 network. [Online]. Available: http://www.internet2.edu/network/.
[27] S. Leinen. What flows in a reserach and education network? [Online]. Available: http://pam2009.kaist.ac.kr/presentation/switch-flows.pdf.
[28] R. Carter and M. Crovella, "Measuring bottleneck link speed in packet switched networks," *Performance Evaluation*, vol. 27 and 28, pp. 297–318, 1996.
[29] V. Paxson, "Measurements and analysis of end-to-end Internet dynamics," Ph.D. dissertation, University of California, Berkeley, Stanford, California, 1997.
[30] K. Lai and M. Baker, "Measuring link bandwidths using a deterministic model of packet delay," in *Proc. of ACM SIGCOMM*, Stockholm, Sweden, 2000, pp. 283–294.
[31] C. Dovrolis, P. Ramanathan, and D. Moore, "Packet dispersion techniques and capacity estimation," *IEEE/ACM Transactions on Networking*, vol. 12, no. 6, pp. 963–977, 2004.
[32] K. Salehin and R. Rojas-Cessa, "Active scheme to measure throughput of wireless access link in hybrid wired-wireless network," *IEEE Wireless Communications Letters*, vol. 1, no. 6, pp. 645–648, 2012.
[33] ——, "Scheme to measure relative clock skew of two Internet hosts based on end-link capacity," *IET Electronics Letters*, vol. 48, no. 20, pp. 1282–1284, 2012.
[34] ——, "Packet-pair sizing for controlling packet dispersion on wired heterogeneous networks," in *Proc. of IEEE ICNC*, CA, USA, 2013, pp. 1031–1035.
[35] S. Keshav, "A control-theoretic approach to flow control," in *Proc. of ACM SIGCOMM*, NY, USA, 1991, pp. 3–15.
[36] N. Hu and P. Steenkiste, "Evaluation and characterization of available bandwidth probing techniques," *IEEE Journal of Selected Areas in Communications*, vol. 21, no. 6, pp. 879–894, 2003.
[37] K. Harfoush, A. Bestavros, and J. Byers, "Measuring bottleneck bandwidth of targeted path segments," *IEEE/ACM Transactions on Networking*, no. 1, pp. 80–92, 2009.
[38] R. Mandeville and J. Perser. RFC 2889 - Benchmarking methodology for LAN switching devices. [Online]. Available: http://www.ietf.org/rfc/rfc2889.txt.
[39] D. Bovet and M. Cesati, *Understanding the Linux Kernel, ch. 5*. Sebastopol, CA, USA: O'Reilly, 2001.
[40] Wireshark. [Online]. Available: http://www.wireshark.org/.
[41] Endace DAG 7.5G2 datasheet. [Online]. Available: http://www.endace.com/assets/files/resources/END_Datasheet_DAG7.5G2_3.0.pdf.
[42] K. Salehin and R. Rojas-Cessa, "Measurement of packet processing time of an Internet host using asynchronous packet capture at the data-link layer," in *Proc. of IEEE ICC*, Budapest, Hungary, 2013, pp. 1143–1147.
[43] A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, "The changing nature of network traffic: Scaling phenomena," *ACM SIGCOMM Computer Communication Review*, vol. 28, no. 2, pp. 5–29, 1998.
[44] Z. Zhang, V. Ribeiro, S. Moon, and C. Diot, "Small-time scaling behaviors of Internet backbone traffic: An empirical study," in *Proc. of IEEE INFOCOM*, CA, USA, 2003, pp. 1826–1836.
[45] R. Govindan and V. Paxson, "Estimating router ICMP generation time," in *Proc. of PAM*, CO, USA, 2002, pp. 1–8.
[46] L. Angrisani, G. Ventre, L. Peluso, and A. Tedesco, "Measurement of processing and queuing delays introduced by an open-source router in a single-hop network," *IEEE Transations on Instrumentation and Measurement*, vol. 55, no. 4, pp. 1065–1076, 2006.
[47] J. Bolot, "End-to-end packet delay and loss behavior in the Internet," in *Proc. of ACM SIGCOMM*, NY, USA, 1993, pp. 289–298.
[48] K. Lai, "Measuring bandwidth," in *Proc. of IEEE INFOCOM*, NY, USA, 1999, pp. 235–245.
[49] R. Kapoor, L. Chen, L. Lao, M. Gerla, and M. Sanadidi, "CapProbe: A simple and accurate capacity estimation technique," in *Proc. of ACM SIGCOMM*, OR, USA, 2004, pp. 67–78.
[50] V. Jacobson. Pathchar - A tool to infer characteristics of Internet paths. [Online]. Available: ftp://ftp.ee.lbl.gov/pathchar/msri-talk.pdf.
[51] B. Mah. pchar: A tool for measuring Internet path characteristics. [Online]. Available: http://www.kitchenlab.org/www/bmah/Software/pchar/.
[52] K. Salehin and R. Rojas-Cessa, "A combined methodology for measurement of available bandwidth and link capacity in wired packet networks," *IET Communications*, vol. 4, no. 2, pp. 240–252, 2010.
[53] InterMapper Web Server. [Online]. Available: https://intermapper.engineering.cenic.org.