Packet-Dispersion Techniques and a Capacity-Estimation Methodology

Constantinos Dovrolis, *Member, IEEE*, Parameswaran Ramanathan, and David Moore, *Member, IEEE*

***Abstract*—The packet-pair technique aims to estimate the capacity of a path (bottleneck bandwidth) from the dispersion of two equal-sized probing packets sent back to back. It has been also argued that the dispersion of longer packet bursts (packet trains) can estimate the available bandwidth of a path. This paper examines such packet-pair and packet-train dispersion techniques in depth. We first demonstrate that, in general, packet-pair bandwidth measurements follow a multimodal distribution and explain the causes of multiple local modes. The path capacity is a local mode, often different than the global mode of this distribution. We illustrate the effects of network load, cross-traffic packet-size variability, and probing packet size on the bandwidth distribution of packet pairs. We then switch to the dispersion of long packet trains. The mean of the packet-train dispersion distribution corresponds to a bandwidth metric that we refer to as average dispersion rate (ADR). We show that the ADR is a lower bound of the capacity and an upper bound of the available bandwidth of a path. Putting all of the pieces together, we present a capacity-estimation methodology that has been implemented in a tool called *pathrate*. We report on our experiences with *pathrate* after having measured hundreds of Internet paths over the last three years.**

***Index Terms*—Available bandwidth, bandwidth estimation, bottleneck link, density estimation, multimodal distributions, network measurements, packet pair, packet train.**

I. Introduction

THE Internet is largely a commercial infrastructure in which users pay for their access to an Internet Service Provider (ISP) and from there to the global Internet. It is often the case that the performance level (and tariff) of these network connections is based on their bit rate, or "network bandwidth," since more bandwidth normally means higher throughput and better quality of service. In such an environment, *bandwidth monitoring* becomes a crucial operation. Users need to check whether they get the access bandwidth that they pay for and whether the network "clouds" that they use are sufficiently provisioned. ISPs also need bandwidth monitoring tools in order to plan their capacity upgrades and to detect congested or underutilized links.

Manuscript received August 2, 2001; revised June 3, 2003; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor M. Crovella. This work was supported by the "Scientific Discovery through Advanced Computing" (SciDAC) program of DOE through Award DE-FC02-01ER25467 and by an equipment donation from Intel Corporation.

C. Dovrolis is with the College of Computing, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: dovrolis@cc.gatech.edu).

P. Ramanathan is with the University of Wisconsin, Madison, WI 53706 USA (e-mail: parmesh@ece.wisc.edu).

D. Moore is with the Cooperative Association for Internet Data Analysis (CAIDA), the University of California at San Diego, San Diego Supercomputer Center, La Jolla, CA 92093-0505 USA (e-mail: dmoore@caida.org).

Digital Object Identifier 10.1109/TNET.2004.838606

Network operators commonly use tools such as MRTG [16] to monitor the utilization of their links with information obtained from the router management software. These techniques are based on counters maintained by routers, and they are normally accurate. Their drawback, however, is that they can only be performed with router access, and such access is usually limited to the corresponding network administrators. Instead, in this paper, we focus on an *end-to-end bandwidth monitoring* approach that requires cooperation of only the path end-points. Even though end-to-end approaches are typically not as accurate as router-based measurements, they often are the only feasible approach for monitoring a path that crosses several networks.

Let us first define two important bandwidth metrics for a network path. Consider a network path $\mathcal{P}$ as a sequence of first-come first-served (FCFS) store-and-forward links that transfer packets from the sender $\mathcal{S}$ to the receiver $\mathcal{R}$. Assume that the path is fixed and unique for the duration of the measurements, i.e., no routing changes or multipath forwarding occur. Each link i transmits data with a constant rate of C_i bits per second, referred to as *link capacity* or *transmission rate*. Two bandwidth metrics that are commonly associated with path $\mathcal{P}$ are the *capacity* C and the *available bandwidth* A. *Capacity* is the minimum transmission rate among all links in $\mathcal{P}$. Note that the capacity does not depend on the traffic load of the path. *Available bandwidth*, on the other hand, is the minimum spare link capacity, i.e., capacity not used by other traffic, among all links in $\mathcal{P}$.

More formally, if H is the number of hops (links) in $\mathcal{P}$, C_i is the capacity of link i, and C_0 is the transmission rate of the sender, then the path capacity is

$$C = \min_{i=0...H} C_i. \tag{1}$$

Additionally, if u_i is the *utilization* of link i (with $0 \leq u_i \leq 1$) over a certain time interval, the average spare capacity of link i is $C_i(1-u_i)$. Thus, the available bandwidth of $\mathcal{P}$ in the same interval can be defined as

$$A = \min_{i=0...H} [C_i(1-u_i)]. \tag{2}$$

Even though the term "available bandwidth" has been previously given various interpretations (such as fair share or bulk TCP throughput), there is growing consensus in the literature for a definition that is equivalent to (2) [10], [15], [24]. A longer discussion of the capacity and available bandwidth metrics, including clarifications for paths with rate limiters, traffic shapers, or time-varying capacity, can be found in [23].

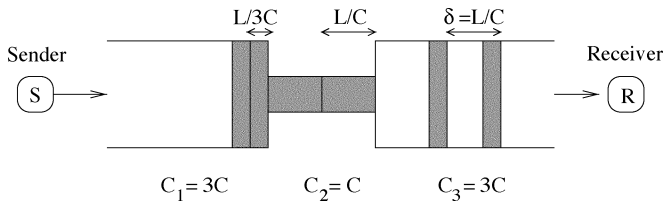

Fig. 1. Graphical illustration of the packet-pair technique. The width of each link corresponds to its capacity. Two packets leave the sender back to back, and they arrive at the receiver with a dispersion that is determined by the narrow link.

The link with the minimum transmission rate determines the capacity, while the link with the minimum spare bandwidth determines the available bandwidth. To avoid the term *bottleneck link*, which has been widely used for both metrics, we refer to the capacity limiting link as *narrow link* and to the available bandwidth limiting link as *tight link*. Note that these two links may be different.

The *packet-pair technique* is a well-known mechanism for measuring the capacity of a path. When a packet is transmitted by a store-and-forward link, it encounters a *transmission (or serialization) delay*, related to the clock rate of the underlying transmission hardware. In a link of capacity C_i, the transmission delay for a packet of size L is $\tau_i = L/C_i$. For now, let us ignore the fact that a packet can carry different encapsulation headers in different links and assume that the packet size L remains constant as the packet traverses the path; we return to this issue in Section III-C. A packet-pair measurement consists of two packets of the same size L sent back to back from $\mathcal{S}$ to $\mathcal{R}$. Without any *cross traffic* in the path, the packet pair will reach $\mathcal{R}$ with a *dispersion* δ (i.e., time spacing from the last bit of the first packet to the last bit of the second packet) equal to $\tau_n \equiv L/C$. Note that τ_n is the transmission delay at the narrow link. The receiver can then estimate the capacity C from the measured dispersion δ, as $C = L/\delta$. Fig. 1 illustrates the packet=pair technique in the case of a three-link path, using the fluid analogy introduced by Jacobson in [8].

Even though simple in principle, the packet-pair technique can produce widely varied measurements and erroneous capacity estimates. The reason is that cross traffic can distort the packet-pair dispersion, leading to capacity underestimation or overestimation. In this paper, we examine the packet-pair bandwidth estimation technique, as well as its generalization to packet trains, in depth. We first demonstrate that, in general, packet-pair bandwidth measurements follow a multimodal distribution and explain what causes multiple local modes. The path capacity is a local mode, often different than the global mode of this distribution, and so it cannot be estimated with standard statistical procedures for the most common value or range. We illustrate the important effects of network load, cross-traffic packet-size variability, and probing packet size on the packet-pair bandwidth distribution. An interesting result is that the conventional wisdom of using maximum-sized packet pairs is not optimal for estimating the capacity of a path. Instead, the probing size of different packet pairs should vary, so that the subcapacity erroneous local modes become wider and weaker.

We then switch to the dispersion of long packet trains. Earlier work [2], [3], [7] assumed that the dispersion of long packet trains is directly related to the available bandwidth of a path.[1] We show that this is not the case. Instead, the mean of the packet-train dispersion distribution corresponds to a bandwidth metric, referred to as average dispersion rate (ADR), that depends in general on the capacity and utilization of *all* links in the path as well as on the routing of cross traffic relative to the measured path. We also show that the ADR is a lower bound of the capacity and an upper bound of the available bandwidth.

Putting all of the pieces together, we present a capacity-estimation methodology that has been implemented in a tool called *pathrate*. *Pathrate* sends many packet pairs to uncover the local modes of the underlying bandwidth distribution and then selects the local mode that corresponds to the capacity. This latter part is also based on the fact that the ADR, measured with long packet trains, is a lower bound for the capacity of the path. We report on our experiences with *pathrate*, after having measured hundreds of Internet paths over the last three years.

The rest of the paper is structured as follows. Section II summarizes previous work on bandwidth estimation. Section III investigates in depth the distribution of packet-pair bandwidth measurements, while Section IV focuses on the distribution of packet-train bandwidth measurements and derives key properties of the ADR. Based on the insight of previous sections, Section V describes the *pathrate* capacity-estimation methodology and tool. We conclude in Section VI.

II. Previous Work

The concept of packet dispersion, as a burst of packets traverses the narrow link of a path, was originally described in [8], and it is closely related to the TCP self-clocking mechanism. However, Jacobson did not consider the effects of cross traffic, and so he did not distinguish between capacity and available bandwidth. Keshav explored the same concept in the context of congestion control [11] and recognized that the dispersion of packet pairs is not related to the available bandwidth when router queues use the First-Come First-Served discipline, and so he focused on fair-queueing instead. Bolot used packet-dispersion measurements to estimate the capacity of a transatlantic link and to characterize the traffic interarrivals [1].

Early work on packet-pair dispersion was followed by more sophisticated variations, focusing on statistical techniques that can extract an accurate capacity estimate from noisy bandwidth measurements. Carter and Crovella created *bprobe*, which uses union and intersection filtering of packet-pair measurements, with variable probing sizes, to produce a final capacity estimate [3]. Lai and Baker, on the other hand, used a kernel density estimator as their statistical filtering tool [12] and maximum-sized probing packets.

Paxson was the first to observe that the distribution of bandwidth measurements is multimodal, and he elaborated on the identification and final selection of a capacity estimate from

[1]Even though [3] did not formalize the available bandwidth definition as in (2), it essentially used the same definition ("the portion of the base bandwidth not used by competing traffic").

these modes [20]. He used both packet pairs and packet trains to estimate the underlying bandwidth distribution. The complete methodology is called "Packet Bunch Modes" (PBM), as Paxson notes in [20] (p. 267):

> "It is unfortunate that PBM has a large heuristic component, as it is more difficult to understand. (...) We hope that the basic ideas underlying PBM—searching for multiple modes and interpreting the ways they overlap in terms of bottleneck changes and multi-channel paths—might be revisited in the future, in an attempt to put them on a more systematic basis."

More recently, the packet-pair technique has been largely revisited, explaining the multiple modes that Paxson observed based on queueing and cross-traffic effects [4], [15], [19]. This paper is an extension of our earlier work [4], however, with some major differences. Specifically, we elaborate on the positive effect of packet pairs with variable sizes (while [4] recommended fixed-size probing packets), define the ADR based on the mean dispersion of packet trains, rather than based on the asymptotic dispersion rate as the train length increases, prove that the ADR is an upper bound for the available bandwidth, and present a significantly different statistical methodology for capacity estimation. [19] proposed a queueing model which shows the effects of cross traffic and identified some key signatures in the packet-pair distribution. Additionally, [19] revealed the negative effect of lower layer headers, and it argued for peak detection as superior to mode detection for capacity estimation. Along a different research thread, [6] showed that it is possible to measure the capacity of targeted path segments using packet-dispersion techniques.

Dispersion techniques using packet trains, instead of packet pairs, have also been proposed for available bandwidth estimation. Carter and Crovella developed a tool called *cprobe* which estimates the available bandwidth from the dispersion of eight-packet trains [3]. Other researchers have proposed that the *ssthresh* variable in TCP's slow-start phase, which should be ideally set to the product of the connection's RTT with the available bandwidth, can be determined from the dispersion of the first three or four ACKs [2], [7]. The underlying assumption in [2], [3], and [7] is that the dispersion of packet trains is inversely proportional to the available bandwidth. However, as it was first shown in [15] and then in [4], this is not the case: the average dispersion of long packet trains is inversely proportional, instead, to the ADR.[2]

More recently, significant progress has been made in the estimation of available bandwidth through end-to-end measurements [10], [15], [24]. The TOPP and SLoPS techniques, of [15] and [10], respectively, use packet streams of variable rates. When the stream rate exceeds the available bandwidth, the stream arrives at the receiver with a lower rate than its rate at the sender. TOPP has the additional advantage that, together with available bandwidth, it can also estimate the capacity of the tight link in the path, and, in some cases, the capacity and available bandwidth of other links in the path. SLoPS, on the other hand, gives more emphasis on the variability of available bandwidth and on the resulting measurement uncertainty (referred to as "grey region"). For a more detailed discussion of the related work in available bandwidth estimation, we refer the reader to [10].

We finally note that packet-dispersion techniques are not the only way to measure bandwidth. A different methodology, called *variable packet size* (VPS) probing, attempts to measure the capacity of *every link in a path*. The underlying idea in VPS probing is based on the variation of one-way delays as the packet size increases [5], [9]. Lai and Baker proposed a variation of this technique, called *packet tailgating*, which avoids the need for ICMP replies from routers [13]. However, the reported capacity estimates with VPS techniques are often inaccurate. A possible explanation is that the presence of layer-two store-and-forward devices (e.g., Ethernet switches) creates additional transmission delays that are not accounted for by these tools [17], [21].

III. Packet-Pair Dispersion

A packet-pair measurement can be described more formally as follows. Consider a network path $\mathcal{P}$ defined by the sequence of link capacities $\mathcal{P} = \{C_0, C_1, \ldots, C_H\}$. Two packets of the same size L are sent from the sender $\mathcal{S}$ to the receiver $\mathcal{R}$. These packets are called *probing packets of size* L. The *dispersion* δ_i of the packet pair after a link i is the time interval between the complete transmission (up to the last bit) of the two packets by link i. The dispersion of the probing packets after the sender $\mathcal{S}$ is $\delta_0 = \tau_0 \equiv L/C_0$, i.e., the two probing packets are sent "back to back." When the packet pair reaches the receiver, $\mathcal{R}$ measures the dispersion δ_H and then computes a bandwidth estimate $b = L/\delta_H$. Since δ_H can vary among different packet-pair measurements, b can be considered a continuous random variable. Suppose that b follows a probability density function (pdf) $\mathcal{B}$ that we refer to as the *packet-pair bandwidth distribution*. Our main objective here is to understand the salient features of the distribution $\mathcal{B}$ and, in particular, those characteristics that relate to the capacity of the path.

First, let us assume that there is no cross traffic in the path. It is easy to see that the dispersion δ_i cannot be lower than the dispersion δ_{i-1} at the previous hop and the transmission delay $\tau_i = L/C_i$ at hop i, i.e., $\delta_i = \max\{\delta_{i-1}, \tau_i\}$. Applying this model recursively from $\mathcal{R}$ back to $\mathcal{S}$, we find that the dispersion at $\mathcal{R}$ is

$$\delta_H = \max_{i=0\ldots H} \tau_i = \frac{L}{\min_{i=0\ldots H}\{C_i\}} = \frac{L}{C} = \frac{L}{C_n} = \tau_n \quad (3)$$

where C_n and τ_n are the capacity and transmission delay at the narrow link, respectively. Consequently, *without cross traffic, all packet-pair bandwidth measurements are equal to the capacity, independent of the probing size* L.

Note that the two probing packets of the same packet pair should have the same size L, so that both packets encounter the same transmission delay at each link. Otherwise, if L_1 is the size of the first probing packet and $L_2 \neq L_1$ is the size of the second probing packet, the dispersion after link i is $\delta_i = \delta_{i-1} + (L_2 - L_1)/(C_i)$ if $\delta_{i-1} > L_1/C_i$, and $\delta_i = L_2/C_i$ otherwise.

[2]Note that [15] used the term "proportional share bandwidth" to describe what we call ADR.

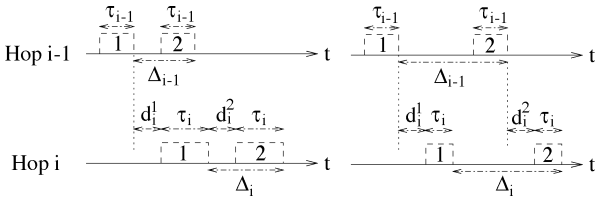

Fig. 2. Two cases of (4).

The dispersion δ_H at the receiver is not always determined by the capacity of the narrow link in that case.[3]

In the presence of cross traffic, probing packets can experience additional queueing delays. Let d_i^1 be the queueing delay of the first probing packet at hop i. Also, let d_i^2 be any additional queueing delay of the second probing packet at hop i, *after the first packet has departed from that link* (see Fig. 2). The dispersion after hop i is

$$\delta_i = \begin{cases} \tau_i + d_i^2, & \text{if } \tau_i + d_i^1 \geq \delta_{i-1} \\ \delta_{i-1} + \left(d_i^2 - d_i^1\right), & \text{otherwise} \end{cases}. \quad (4)$$

Note that, when $\tau_i + d_i^1 < \delta_{i-1}$ and $d_i^2 < d_i^1$, the dispersion from hop $i-1$ to hop i decreases, i.e., $\delta_i < \delta_{i-1}$. This effect can cause a dispersion at $\mathcal{R}$ that is lower than the dispersion at the narrow link ($\delta_H < L/C_n$). Note that this can only happen if there are additional links after the narrow link.[4] We refer to such links as *post-narrow links*. The last observation implies that *the capacity of the path cannot be estimated simply from the minimum measured dispersion* (or, equivalently, the maximum bandwidth measurement) because the minimum dispersion could have resulted at a post-narrow link.

In order to examine the properties of the packet-pair bandwidth distribution $\mathcal{B}$ in a controllable and repeatable manner, we used the network simulator NS. Simulations allow us to investigate the effects of cross traffic in packet-pair dispersion, avoiding issues such as route changes, multichannel links, timestamping accuracy, and resolution, that can distort the measurements. Together with simulations, we also present measurements from Internet paths.

The simulated model follows the path description given earlier, i.e., $\mathcal{S}$ sends packet pairs to $\mathcal{R}$, and the latter computes bandwidth estimates b from the measured dispersions δ_H. The bandwidth distribution $\mathcal{B}$ is estimated from a histogram of 1000 packet-pair measurements. The Appendix describes in detail the statistical procedure that we use to detect *local modes* in $\mathcal{B}$, as well as a heuristic for selecting the *bin width* ω. In summary, a local mode estimates the range of a local maximum in $\mathcal{B}$. The *strength* of a local mode is the number of measurements in the central bin of that mode, i.e., the maximum number of measurements in a range of width ω in that mode. The *width* of a local mode, on the other hand, is related to the range in which that mode extends. The *global mode* of $\mathcal{B}$ is the local mode with the maximum strength, i.e., the most common range of bandwidth measurements with width ω. For example, Fig. 12 shows a distribution with three local modes; the global mode has a strength S_2 and a width W_2.

[3]Variations of the packet-pair technique, with probing packets of different sizes, have been proposed for different estimation purposes [6], [13], [19].

[4]If there is more than one link with capacity C, the narrow link is the last of them in the path.

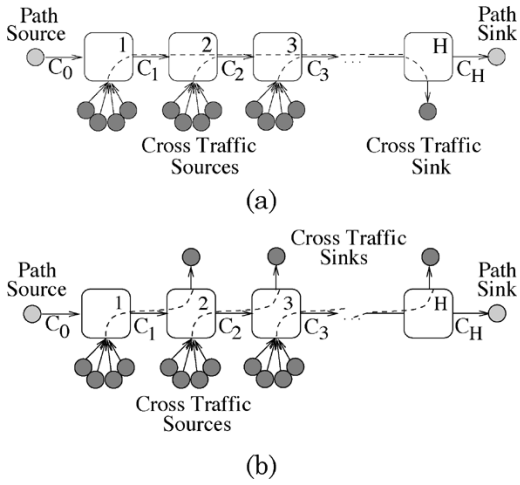

Fig. 3. Two cases of cross-traffic routing. (a) Path-persistent cross traffic. (b) One-hop (persistent cross traffic.

Unless noted otherwise, the probing size L is fixed at 1500 bytes. Cross traffic is generated from 16 sources at each hop with Pareto interarrivals and shape parameter $\alpha = 1.9$, i.e., the packet interarrivals are heavy-tailed. The cross-traffic packet size is denoted by L_c, and it is either constant or it is uniformly distributed in the [40, 1500] (bytes) range.

The routing of cross-traffic packets relative to the measured path is also important. Fig. 3 shows two extreme cases. In Fig. 3(a), cross traffic follows the same path with the packet pairs (*path-persistent cross traffic*). In Fig. 3(b), cross traffic exits the path after one hop (*one-hop persistent cross traffic*). We explain the importance of cross-traffic routing in Section IV. For now, we note that the following simulations use one-hop persistent cross traffic.

A. Effect of Network Load

Let us first examine *the effect of network load on the packet-pair bandwidth distribution* $\mathcal{B}$. Fig. 4 shows the histogram of $\mathcal{B}$ for a path $\mathcal{P} = \{100, 75, 55, 40, 60, 80\}$ (all capacities in Mb/s), with a bin width of 2 Mb/s. Note that the path capacity is $C = 40$ Mb/s, while the post-narrow links have capacities of 60 and 80 Mb/s, respectively. In Fig. 4(a), all links are 20% utilized, whereas in Fig. 4(b) all links are 80% utilized.

When the path is lightly loaded ($u = 20\%$), the capacity value of 40 Mb/s is prevalent in $\mathcal{B}$, and it forms the capacity mode (CM), that is, the global mode of the distribution in this case. CM is formed by packet pairs that did not get queued behind cross-traffic packets in the path. Bandwidth measurements at the left of CM are caused by cross-traffic packets that interfere with packet pairs, increasing their dispersion, and causing a subcapacity dispersion range (SCDR). For instance, the SCDR in Fig. 4(a) is between 10–40 Mb/s. The reason we observe several local modes in the SCDR is discussed later in this section.

Bandwidth measurements at the right of CM are caused at the post-narrow links when the first probing packet is delayed more than the second. The corresponding local modes are referred

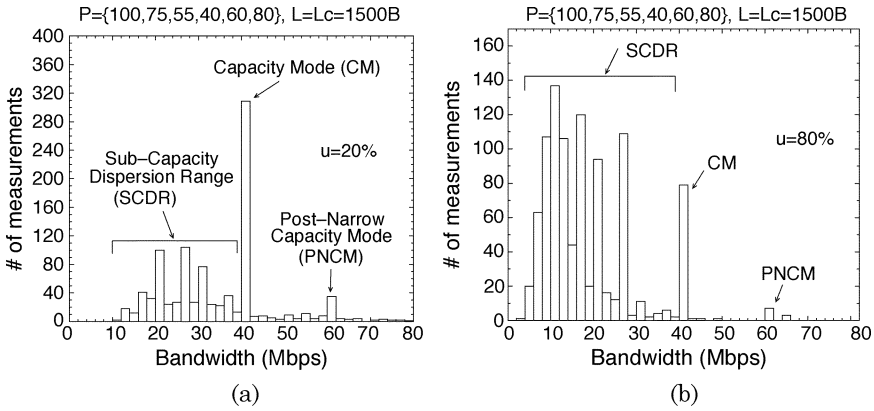

Fig. 4. Effect of network load on the packet-pair bandwidth distribution $\mathcal{B}$ (simulations). (a) Light load conditions. (b) Heavy load conditions.

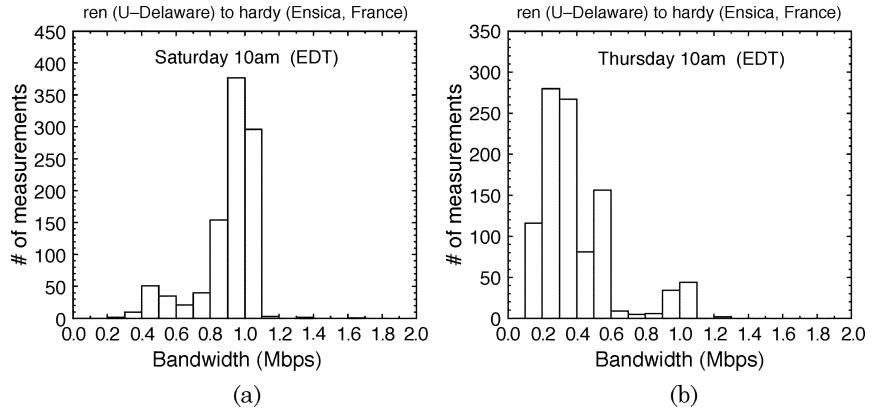

Fig. 5. Effect of network load on the packet-pair bandwidth distribution $\mathcal{B}$ (measurements). (a) Light load conditions. (b) Heavy load conditions.

to as post-narrow capacity modes (PNCMs). Note a PNCM at 60 Mb/s, which is the capacity of the link just after the narrow link. That local mode is created when the first probing packet is delayed long enough for a packet pair to be serviced back to back in that link. A link with capacity C_i can form a PNCM if all links later in the path have higher capacities, i.e., if $C_i < C_j$ for any link j with $i < j \leq H$.

In the case of heavy load $(u = 80\%)$, the probability of cross-traffic packets interfering with the probing packets becomes much larger, and *CM is not the global mode of* $\mathcal{B}$ [see Fig. 4(b)]. Instead, the global mode resides in the SCDR, which now dominates the bandwidth measurements. In fact, in heavily congested paths, CM may not even appear as a local mode if almost every packet pair is affected by cross traffic.

A key point to take from Fig. 4(b) is that *the path capacity cannot be estimated, in the general case, by statistical techniques that extract the most common bandwidth value or range*. Instead, we must analyze the queueing effects that cause different local modes, understand what distinguishes CM from SCDR and PNCM modes, and choose a probing size L that will make CM relatively stronger than those erratic modes.

Fig. 5 shows packet-pair bandwidth distributions based on Internet measurements, rather than simulations. The measured path is from the University of Delaware to Ensica in Toulouse, France. The capacity of the path is 1 Mb/s, limited by Ensica's access link. The distribution of Fig. 5(a) resulted from measurements on a Saturday morning in Delaware (afternoon in France), while the distribution of Fig. 5(b) resulted from measurements on a Thursday morning in Delaware. Even though we do not know the actual utilization at each link of the path, it is reasonable to expect that the path is much more loaded on a weekday than on a Saturday. The impact of the network load is obvious in Fig. 5. *When the path is lightly loaded, CM is prevalent and thus easy to measure. In heavier loads, on the other hand, the SCDR is prevalent, and CM is only a minor local mode*. There are no significant PNCMs in these measurements.

B. Effect of Cross-Traffic Packet-Size Variability

Let us now investigate *what causes local SCDR modes*. Fig. 6 shows $\mathcal{B}$ for the same path as in Fig. 4, when the cross-traffic packet size L_c is fixed to 1500 bytes [Fig. 6(a)] and when it varies uniformly in the range [40, 1500] bytes [Fig. 6(b)]. In the first case, the probing size L is also 1500 bytes, while in the second case it is set to 770 bytes, i.e., the average of the cross-traffic packet range [40, 1500].

When all packets in the path have the same size $(L_c = L)$, it is simple to explain the local modes in the SCDR of Fig. 6(a). The basic idea is that *SCDR local modes are created when a number of cross-traffic packets interferes with the packet pair in specific links of the path*. For instance, consider the path $\mathcal{P} = \{100, 60, 40\}$. A local mode at 30 Mb/s can be caused when a single cross traffic packet interferes between packet pairs at the 60-Mb/s link. In that case, the packet-pair dispersion after the narrow link is $\delta_n = (L/40) + (2(L/60) - (L/40)) = (L/30)$

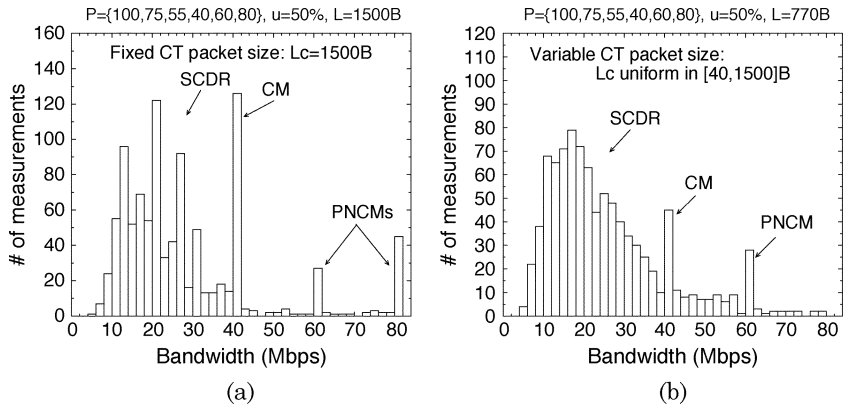

Fig. 6. Effect of the cross-traffic packet size L_c on the packet-pair bandwidth distribution $\mathcal{B}$. (a) Fixed cross-traffic packet size. (b) Variable cross-traffic packet size.

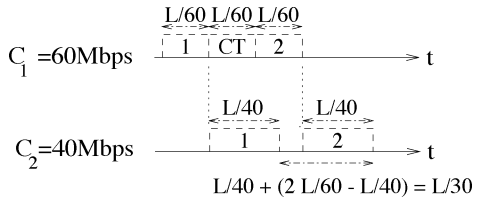

Fig. 7. Generation of a 30-Mb/s local mode in a path $\mathcal{P} = \{100, 60, 40\}$.

(see Fig. 7). Similarly, a mode at 20 Mb/s can be caused by a single packet interfering between packet pairs at the 40-Mb/s link, or by two packets interfering at the 60-Mb/s link, and so on.

On the other hand, when cross-traffic packet sizes vary uniformly in the range [40, 1500] [Fig. 6(b)], the resulting packet-pair dispersion is much less predictable. If the cross-traffic packet-size distribution has no local modes (such as the uniform distribution), the SCDR in $\mathcal{B}$ has no local modes either. On the other hand, the CM and some of the PNCMs are still distinct modes in the distribution of Fig. 6(b), because they are caused by probing packets serviced back to back at the narrow or post-narrow links, respectively, independent of the cross-traffic packet-size distribution.

Several measurement studies have shown that the packet-size distribution in the Internet has strong modalities, centered around three or four common values [14], [26]. Specifically, about 50% of the packets are 40 (or 44) bytes, 20% are 576 bytes, and 15% are 1500 bytes. These dominant packet sizes would cause a packet-pair bandwidth distribution that is more similar to the discrete dispersion effects of Fig. 6(a), rather than the continuous dispersion effects of Fig. 6(b). Thus, *when measuring real Internet paths with fixed-size probing packets, we should expect strong SCDR local modes*.

C. Effect of the Probing Packet Size

We now focus on the effect of the probing size L on the packet-pair bandwidth distribution $\mathcal{B}$. The conventional wisdom, as reflected, for instance, in [20] or [12], is that the best value of L is the maximum nonfragmented packet size, i.e., the path Maximum Transmission Unit (MTU) size. The reason is that a larger L leads to wider dispersion that is easier to measure, more robust to queueing delay noise, and less sensitive to the timestamping resolution at the receiver. Here, we first investigate the effect of the probing size on the strength of the PNCM, CM, and SCDR modes, showing that MTU-sized probing packets are not optimal for capacity estimation. We also show that, if the probing size L varies among different packet pairs, the SCDR modes become wider and weaker, making the CM mode relatively stronger.

Let us first examine *the effect of a fixed probing size L on the strength of the PNCM, CM, and SCDR modes*. Suppose that a packet pair arrives at a link i of capacity C_i. If a cross-traffic packet arrives at link i in the time interval between the arrival of the first and second probing packets, which is of duration L/C_i, it will interfere with the probing packets, increasing the dispersion δ_i above τ_i. *The larger L is, the higher the probability of an interfering cross-traffic arrival, and the more prevalent the SCDR will be in $\mathcal{B}$.*

The effect of a fixed probing size is clear in Fig. 8, where $\mathcal{B}$ is shown for a small probing size (L = 100 B), and for a large probing size (L =1500 B). In both cases, the path configuration, load, and cross-traffic packet-size distribution are the same. Note that the probing size L is qualified as "small" or "large," relative to the size of cross-traffic packets. The SCDR in the case of small L is much weaker than in the case of large L, because it is less likely for cross-traffic packets to interfere between small probing packets.

As L decreases, on the other hand, the dispersion decreases proportionally, and it becomes more susceptible to distortion at post-narrow links. Suppose that L = 100 B, $\mathcal{P}$ ={40, 80}, and that a packet pair leaves the 40-Mb/s narrow link back to back, i.e., with $\delta_0 = 20$ μs. It only takes at least 100 bytes of cross traffic interfering at the 80-Mb/s link to make the packet pair depart from that link back to back, i.e., with $\delta_1 = 10$ μs. In general, *as L decreases, the formation of PNCMs becomes more likely*. This is shown in Fig. 8. Note the PNCM at 60 Mb/s in Fig. 8(a), which is actually stronger than the CM located at 40 Mb/s. On the other hand, there are no significant PNCMs when L = 1500 B, as shown in Fig. 8(b).

We next consider the following question: *in order to estimate the capacity of a path, is it better to use the same probing size L in all packet pairs, or a variable L across different packet pairs*? The key idea here is that, if L is the same in all packet pairs, SCDR local modes will be created when cross-traffic packets of certain common sizes interfere between packet pairs. For example, if L = 1500 B, the interference of a 1500-byte cross-

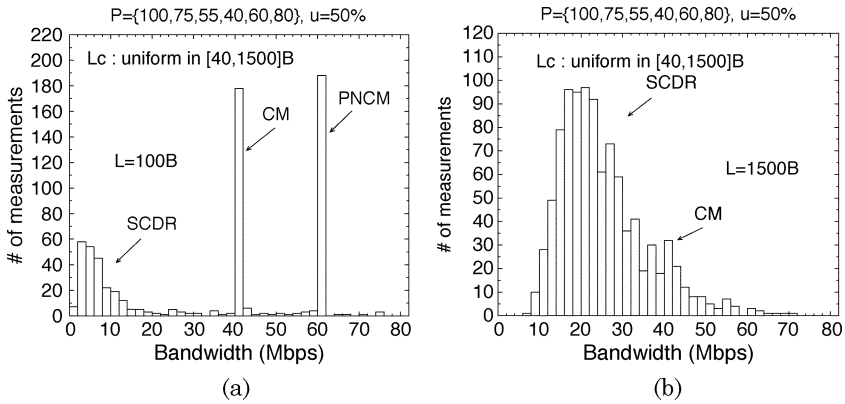

Fig. 8. Small versus large probing size L. (a) $L = 100$ B. (b) $L = 1500$ B.

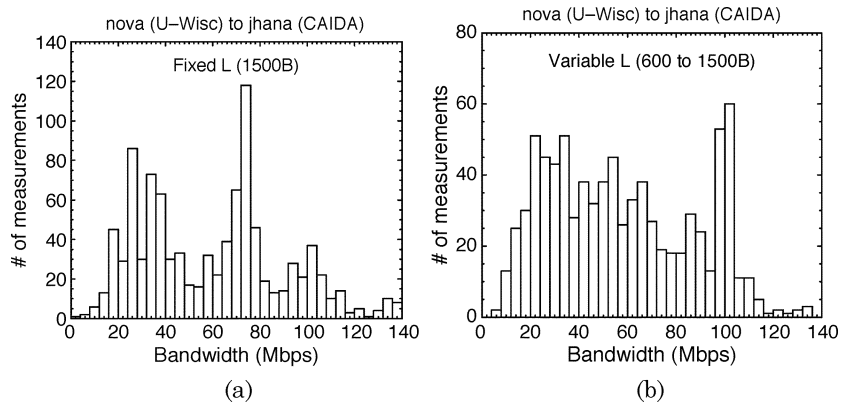

Fig. 9. Fixed versus variable-sized packet-pair bandwidth measurements. (a) $L = 1500$ B. (b) L: 600–1500 B.

traffic packet at a 40-Mb/s link can cause an SCDR mode at 20 Mb/s. As noted in Section III-B, there are three to four common packet sizes in Internet traffic, and so the creation of strong SCDR modes should be expected when all packet pairs have the same probing size L (see also Fig. 6).

Suppose, instead, that L varies uniformly across different packet-pair experiments between 50–1500 B. The interference of 552-byte cross-traffic packets between a packet pair at a 40-Mb/s link will generate bandwidth measurements uniformly distributed between 3.3–29 Mb/s, depending on the probing size L. So, *the use of variable-sized probing packets distributes the subcapacity measurements in a larger bandwidth range, making the SCDR modes wider and thus weaker compared to CM*.

The difference between fixed and variable probing packet sizes is clear in Fig. 9. The two bandwidth distributions resulted from 1000 packet-pair experiments in the path from University of Wisconsin to CAIDA in San Diego, CA. In the first case [Fig. 9(a)], all packet pairs consist of probing packets with size $L = 1500$ B. Note the presence of strong SCDR modes around 72, 35, and 26 Mb/s.[5] CM, located around 100 Mb/s, appears only as a weak mode compared to these SCDR modes. In Fig. 9(b), on the other hand, the probing size varies uniformly between 600–1500 bytes. Notice that the subcapacity measurements have been spread throughout a wider range and that the SCDR modes are weaker compared to Fig. 9(a). The CM has accumulated roughly the same number of measurements in both Figs. 9(a) and (b), but it is easier to detect in the latter, because it is stronger compared to the SCDR modes.

Given all previous constraints, what should the probing size L be? We showed that using a range of L, instead of a certain value, spreads the subcapacity measurements making the SCDR modes wider and thus weaker. We also showed that using probing packets that are too small compared to cross-traffic packets can create intense PNCMs, while using probing packets that are too large can lead to a prevalent SCDR and a weak CM. Combining these effects, we see that *a good compromise for the probing packet size is to use a range of sizes, say from $L_{\min}$ to $L_{\max}$, where $L_{\min}$ is not a very small packet and $L_{\max}$ is not a very large packet, compared to cross traffic packets*.

We need to consider two additional practical constraints on $L_{\min}$. First, encapsulation headers in lower layers of the protocol hierarchy can cause a significant capacity underestimation when using small probing packets. Suppose that a link has a capacity C_{L2} at the link layer (layer-two). At the IP layer (layer-three), the link will deliver a lower rate than its nominal transmission rate C_{L2}, due to the overhead of layer-two headers [19]. The transmission latency for an IP packet of size L_{L3} bytes is

$$\delta_{L3} = \frac{L_{L3} + H_{L2}}{C_{L2}} \tag{5}$$

[5]The estimation of local modes is performed using the statistical procedure described in the Appendix and not based on visual inspection of these histograms.

where H_{L2} is the size of layer-two headers that encapsulate the IP packet. So, the capacity C_{L3} of the link at layer-three is

$$C_{L3} = \frac{L_{L3}}{\delta_{L3}} = C_{L2}\frac{L_{L3}}{L_{L3}+H_{L2}} = C_{L2}\frac{1}{1+\frac{H_{L2}}{L_{L3}}}. \quad (6)$$

Note that the layer-three capacity depends on the size of the IP packet relatively to the layer-two header size. To reduce the effect of layer-two headers on the measured layer-three capacity, we need to use probing packets that are significantly larger than the typical layer-two encapsulation headers (5–50 bytes). In *pathrate*, for instance, the minimum probing size $L_{\min}$ is set to 550 bytes.

Second, we need to consider the per-packet processing time at the receiver. An end-host can only measure the dispersion of a packet pair when the latter is larger than a certain lower bound δ_m. This dispersion δ_m is determined by the latency to receive a packet from the network interface, process the packet at the kernel protocol stack, move the packet from kernel to user space through a *recvfrom* system call, timestamp the arrival, and so on, before waiting for the second probing packet. Given δ_m for a particular host, the maximum possible capacity that can be measured for a packet size L is L/δ_m. For example, with $\delta_m = 10$ μs and $L = 800$ B, the maximum capacity that can be measured is 640 Mb/s. On the other hand, when a rough estimate $\bar{C}$ of the capacity is known, the minimum packet size should be set to $L_{\min} > \bar{C}\delta_m$.

D. Summary of Packet-Pair Dispersion

The packet-pair bandwidth distribution $\mathcal{B}$ can be viewed as a sequence of local modes imposed on an underlying random measurement noise. In general, the path capacity cannot be estimated from the most common measurement (global mode) or from the maximum bandwidth measurement. Each local mode in $\mathcal{B}$ is caused by a commonly occurring queueing event that leads to a specific dispersion range. Some local modes are below the capacity (SCDR), some above the capacity (PNCMs), and one of them (CM) is normally the capacity. CM is typically the global mode of $\mathcal{B}$ under light load conditions. The SCDR accumulates more measurements as the load increases however, and it exhibits several local modes if the cross-traffic packet-size distribution is multimodal. Regarding the probing packet size, using a range of values, instead of a fixed size, spreads the subcapacity measurements making the SCDR modes wider and, thus, weaker. Finally, using probing packets that are too small compared to cross-traffic packets can create intense PNCMs, while using probing packets that are too large can cause a prevalent SCDR and a weak CM.

IV. Packet-Train Dispersion

As a generalization of the packet-pair technique, $\mathcal{S}$ can send N back-to-back packets of size L to $\mathcal{R}$ with $N > 2$; we refer to these probing packets as a *packet train* or simply train of length N. $\mathcal{R}$ measures the *total dispersion* $\Delta(N) = \sum_{k=1}^{N-1}\delta^k$ of the train, from the first to the last packet, where δ^k is the dispersion between packets k and $k+1$. From $\Delta(N)$, $\mathcal{R}$ calculates a bandwidth measurement $b(N)$ as

$$b(N) = \frac{(N-1)L}{\Delta(N)}. \quad (7)$$

$b(N)$ can also be written as $b(N) = L/\bar{\delta}(N)$, where

$$\bar{\delta}(N) = \frac{\sum_{k=1}^{N-1}\delta^k}{N-1} \quad (8)$$

is the average dispersion among successive packet pairs of the train.

Without cross traffic in the path, all bandwidth measurements $b(N)$ will be equal to the capacity, just as in the packet-pair case. So, the capacity of an empty path can be measured with packet trains of any length. One should use packet trains, however, when the narrow link is multichanneled [20]. In a p-channel link of total capacity C, the individual channels forward packets in parallel at a rate of C/p, and so the link capacity can be measured from the dispersion of packet trains with $N = p + 1$.

In a nonempty path, $\Delta(N)$ will vary across different train measurements. In that case, $b(N)$ can be considered a continuous random variable, and we refer to its pdf as the *packet-train bandwidth distribution* $\mathcal{B}(N)$. It may seem at first that using long packet trains simplifies the problem of capacity estimation compared to packet pairs. One can argue such a hypothesis because trains lead to larger dispersion values, which would be less sensitive to measurement noise. However, as we show in this section, this is not the case. The key idea is that, although the dispersion $\Delta(N)$ becomes larger as N increases, so does the cross-traffic noise that is introduced in the dispersion $\Delta(N)$. In other words, *as the train length N increases, more cross-traffic packets can interfere with the packet train, resulting in bandwidth measurements that are less than the path capacity C.*

A. Packet-Train Bandwidth Distribution

We first make three observations, based on simulation and experimental results, on the relation between N and $\mathcal{B}(N)$. Fig. 10 shows histograms of $\mathcal{B}(N)$ for four increasing values of N at a simulated path $\mathcal{P} = \{100, 75, 55, 40, 60, 80\}$ with $u = 80\%$ in all links. Fig. 11 shows histograms of $\mathcal{B}(N)$ for four increasing values of N at an Internet path from the University of Wisconsin to the University of Delaware. All histograms are based on 1000 train measurements.

The first observation is that, *as N increases, the CM and PNCMs become weaker, until they disappear, and the SCDR prevails in $\mathcal{B}(N)$*. This is because, as N increases, the probability that a packet train will encounter additional dispersion due to cross-traffic packets increases as well. When the train length is sufficiently large, almost every bandwidth measurement is less than the path capacity due to cross-traffic interference. This observation also implies that *the optimal train length for generating a strong CM is $N = 2$*, i.e., to use packet pairs. Longer packet trains are more likely to cause capacity underestimation.

A second observation is that *the variability of $b(N)$ decreases as N increases*. In Figs. 10 and 11, this is shown both by the reduced distribution range and by the suppression of local modes as the train length increases. This observation can be explained

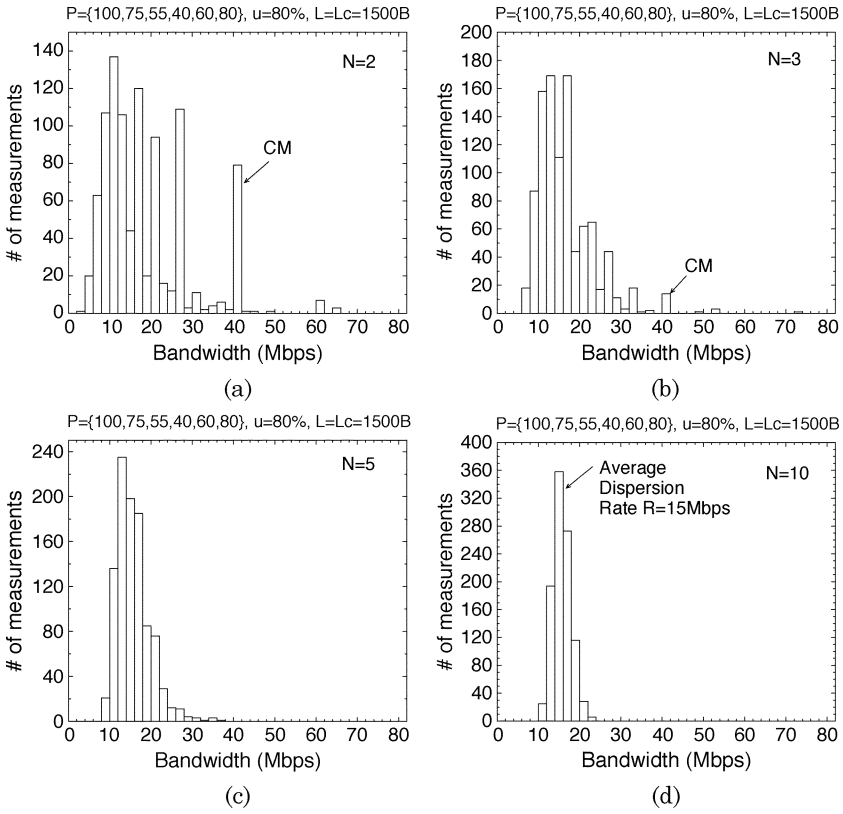

Fig. 10. Effect of the packet train length N (simulations). (a) Packet pairs ($N = 2$). (b) Packet trains with $N = 3$. (c) Packet trains with $N = 5$. (d) Packet trains with $N = 10$.

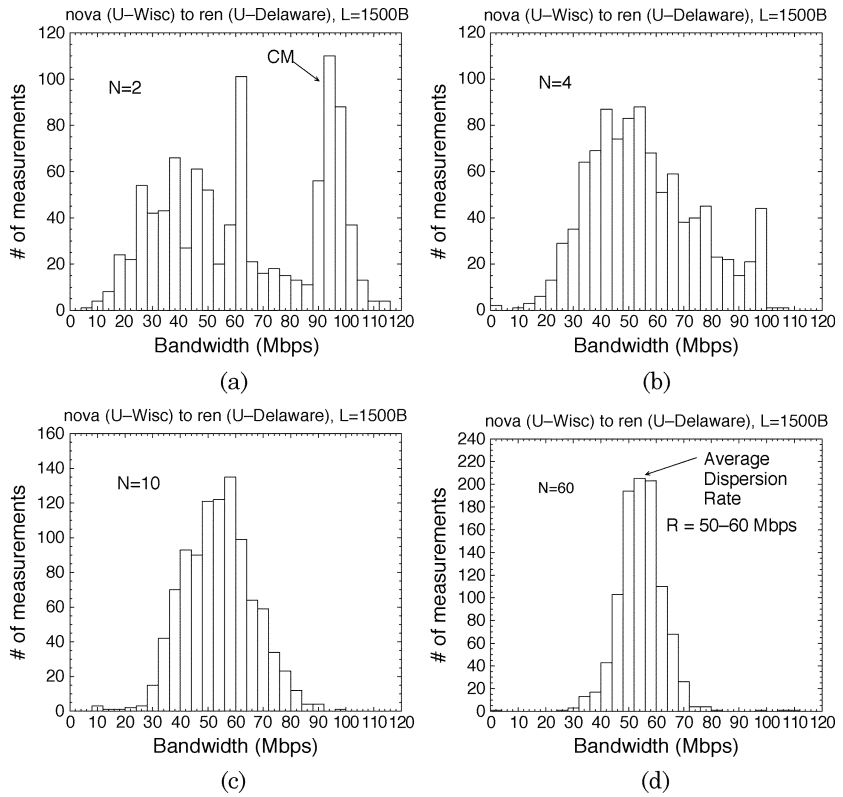

Fig. 11. Effect of the packet train length N (measurements). (a) Packet pairs ($N = 2$). (b) Packet trains with $N = 4$. (c) Packet trains with $N = 10$. (d) Packet trains with $N = 60$.

from (8) as follows: if σ_δ^2 is the variance of the packet-pair dispersion δ, and assuming that the packet-pair dispersions δ^k ($k = 1 \ldots N-1$) are independent, the variance of $\bar{\delta}(N)$ is $\sigma_\delta^2/(N-1)$. Since the train bandwidth $b(N)$ is uniquely determined by $\bar{\delta}(N)$, the variance of $b(N)$ decreases as N increases.

A third observation is that, *if N is sufficiently large, the location of the packet train bandwidth distribution $\mathcal{B}(N)$ becomes independent of N*. In other words, as the train length increases, the resulting bandwidth measurements tend toward a certain value that we refer to as ADR. We next use a simple model of packet train dispersion, based on a fluid cross-traffic model, to explain the effect of N on the distribution $\mathcal{B}(N)$, derive the ADR in a multihop path, and show the relation between the capacity, available bandwidth, and ADR.

B. Average Dispersion Rate

Consider a path $\mathcal{P} = \{C_0, C_1, \ldots C_H\}$ from $\mathcal{S}$ to $\mathcal{R}$. At link i, the average cross-traffic rate is S_i ($S_i < C_i$), the available bandwidth is $A_i = C_i - S_i$, and the utilization is $u_i = S_i/C_i$. We assume that links use the FCFS queueing discipline and that they are adequately buffered to avoid any packet losses. $\mathcal{S}$ sends trains of length N to $\mathcal{R}$ at the source rate C_0. The probing packet size is L bytes. Each train is sent after the previous train has been received, and so different trains do not interact while in transit. The train dispersion $\Delta_i(N)$ after link i is the time interval between the complete transmission of the first and the last packets of the train from link i. The initial dispersion of the train after the source is $\Delta_0(N) = L(N-1)/C_0$, i.e., the probing packets are sent back to back. The train arrives at $\mathcal{R}$ with a dispersion $\Delta(N) \equiv \Delta_H(N)$, resulting in a bandwidth measurement $b(N) = (N-1)L/\Delta(N)$. We are interested in the *bandwidth metric R that corresponds to the mean packet train dispersion* $E[\Delta(N)]$, i.e.,

$$R \equiv \frac{(N-1)L}{E[\Delta(N)]}. \tag{9}$$

We refer to R as the ADR.

1) ADR in a Single-Hop Path: Let us start with a single-hop path $\mathcal{P} = \{C_0, C_1\}$, where $C_0 \geq A_1$. We assume that the cross traffic follows a fluid model, and so the amount of cross traffic that arrives at link i in any time interval T is $S_i T$. So, the cross traffic that arrives in the link during a train's initial dispersion $\Delta_0(N)$ is $X_1 = S_1 \Delta_0(N)$. The traffic X_1 is interleaved with the packet train at the FCFS queue of the link. The dispersion of the packet train after the link is

$$\Delta_1(N) = \frac{(N-1)L + X_1}{C_1} = \frac{(N-1)L}{C_1}\left(1 + u_1 \frac{C_1}{C_0}\right) \tag{10}$$

and so the ADR is

$$R = \frac{(N-1)L}{E[\Delta_1(N)]} = \frac{C_1}{1 + u_1 \frac{C_1}{C_0}} \leq C_1. \tag{11}$$

Note that *R does not depend on the train length N*. This agrees with the last observation of Section IV-A, according to which the ADR is independent of the train length N.

Is (11) valid for any train length N, however? To answer this question, we need to depart from the previous fluid model and take into account that cross traffic appears as a random process of distinct packet arrivals. In that case, the amount of cross traffic X_1 that arrives at the link during a time interval $\Delta_0(N)$ may be different than the mean $E[X_1] = S_1 \Delta_0(N)$. As N increases, however, we expect that the cross-traffic rate $X_1/\Delta_0(N)$ during the arrival of a particular packet train of length N will converge to the average rate S_1, and so we can approximate X_1 by its mean $E[X_1]$. Thus, (11) can be considered a good approximation for the ADR with packetized cross traffic as long as N is sufficiently large. As shown in Fig. 11(d), which resulted from experiments with real Internet traffic, a train length of a few tens of packets is typically sufficient to approximate the location of the distribution $\mathcal{B}(N)$.

Equation (11) also shows that, for capacity-estimation purposes, it is better to inject probing packets in the path from a higher bandwidth interface (higher C_0), since the cross-traffic error term $u_1\, C_1/C_0$ is then reduced. For instance, suppose that we measure the capacity of a path using packet trains and that $C_0 = 100$ Mb/s and $C_1 = 100$ Kb/s. Even if the narrow link is almost saturated ($u \approx 1$), the term $u_1\, C_1/C_0$ will introduce an underestimation error that is less than 0.1%. This also explains why it is harder to estimate the capacity of paths in which the narrow link capacity is of the same order of magnitude with the source transmission rate.

2) ADR in a Multihop Path With One-Hop Persistent Cross Traffic: Consider now the general case of a multihop path $\mathcal{P} = \{C_0, C_1, \ldots C_H\}$. We assume again that the cross traffic follows the fluid model, and so the amount of cross traffic arriving at link i in any time interval T is $S_i T$. Additionally, we assume that the cross traffic is one-hop persistent [see Fig. 3(b)]. This assumption guarantees that the amount of interfering cross traffic at link i does not depend on the amount of interfering cross traffic at previous links. Let R_i be the ADR at the output of link i, with $R_0 = C_0$. We derive R_i as a function of R_{i-1} for $i = 1, \ldots H$.

If $R_{i-1} < A_i$, we have that $R_{i-1} + S_i < C_i$, and so cross traffic interferes between probing packets without increasing their dispersion. Thus, the ADR at the output of link i is $R_i = R_{i-1}$.

If $R_{i-1} \geq A_i$, we can follow the same derivations that resulted in (11) to show that

$$R_i = \frac{C_i}{1 + u_i \frac{C_i}{R_{i-1}}} = \frac{R_{i-1} C_i}{S_i + R_{i-1}} = \frac{R_{i-1} C_i}{(C_i - A_i) + R_{i-1}}. \tag{12}$$

Putting the previous two cases together, we have that

$$R_i = \begin{cases} R_{i-1} \frac{C_i}{S_i + R_{i-1}}, & \text{if } R_{i-1} \geq A_i \\ R_{i-1}, & \text{otherwise.} \end{cases} \tag{13}$$

So, if we know the capacity and available bandwidth at each link of the path, we can derive the ADR applying (13) recursively from $i = 1$ to $i = H$, assuming that the cross traffic is one-hop persistent. Note that the ADR is determined, in general, by the capacity and available bandwidth *of each link in the path*. This is fundamentally different than the end-to-end available bandwidth $A = \min A_i$, which depends on the utilization and capacity of only the tight link.

C. ADR and Cross-Traffic Routing

If the cross traffic is not one-hop persistent, the interfering cross traffic at link i depends on the amount of cross traffic at previous links. To illustrate this point, consider the two-hop path $\mathcal{P} = \{C_0, C_1, C_2\}$ with $C_0 \geq C_1 \geq C_2$. We compare the ADR between one-hop persistent and path persistent cross traffic (see Fig. 3), under the same load conditions.

Suppose that the average cross-traffic rate is $S_1 = r_1 < C_1$ at link-1, and $S_2 = r_1 + r_2 < C_2$ at link-2, with $r_1, r_2 > 0$. In the case of one-hop persistent cross traffic, we can use (13) to show that the ADR after each link is

$$R_1^{1p} = \frac{C_1}{1 + \frac{r_1}{C_0}} \quad R_2^{1p} = \frac{C_2}{1 + \left(1 + \frac{r_1}{C_0}\right)\left(\frac{r_1}{C_1} + \frac{r_2}{C_1}\right)} \tag{14}$$

because $C_0 \geq A_1$ and $R_1^{1p} \geq A_2 = C_2 - S_2$.

In the path-persistent model, the cross-traffic rate entering link-1 is r_1 and the cross traffic entering link-2 is r_2. So, the aggregate load at link-2 is $S_2 = r_1 + r_2$, as in the one-hop persistent case. The ADR at the exit of link-1 is the same as before:

$$R_1^{pp} = R_1^{1p} = \frac{C_1}{1 + \frac{r_1}{C_0}}. \tag{15}$$

The train dispersion at the exit of link-1 is

$$\Delta_1 = \frac{(N-1)L + r_1\Delta_0}{C_1} \tag{16}$$

with $\Delta_0 = (N-1)L/C_0$. The additional cross traffic that interferes with the train at link-2 is $X_2 = \Delta_1 r_2$, and so the train dispersion at the exit of link-2 is

$$\Delta_2 = \frac{(N-1)L + r_1\Delta_0 + r_2\Delta_1}{C_2}. \tag{17}$$

So, the ADR after link-2 is

$$R_2^{pp} = \frac{(N-1)L}{\Delta_2} = \frac{C_2}{\left(1 + \frac{r_1}{C_0}\right)\left(1 + \frac{r_2}{C_1}\right)}. \tag{18}$$

It is easy to show that $R_2^{pp} > R_2^{1p}$, i.e., *the one-hop persistent model results in lower ADR*. The two models produce different results because, in the path-persistent model, the cross traffic that interferes with the train at link-1 has already been "spaced out" after link-1 to a rate that is lower than r_1. So, the cross traffic that interferes with the train at link-2 has a lower rate than $r_1 + r_2$.

D. Relation Between ADR, Capacity, and Available Bandwidth

It is easy to show that that the ADR is a lower bound for the capacity C of a path.

Proposition 1: The ADR at the receiver is $R \leq C$.

Proof: According to (12), if $R_{i-1} \geq A_i$, then $R_i \leq C_i$. So, from (13), the ADR after each link i is either less than the available bandwidth A_i, and thus less than the capacity C_i, or it becomes less than the capacity C_i. If the narrow link is link n, this means that $R_n \leq C_n = C$. Since the ADR after link i is never larger than the dispersion rate after link $i-1$, we have that $R \equiv R_H \leq C$. Q.E.D.

We next prove that the ADR is an upper bound for the available bandwidth A of a path, as long as the source rate C_0 is not lower than A.

Proposition 2: If $C_0 \geq A$, the ADR at the receiver is $R \geq A$.

Proof: Consider first a link i in which $A_i \leq R_{i-1}$. From (12), we can check whether $R_i \geq A_i$ by examining the inequality $A_i^2 - (C_i + R_{i-1})A_i + C_i R_{i-1} \geq 0$. This inequality holds because $A_i \leq R_{i-1}$ and $A_i \leq C_i$. So, if $A_i \leq R_{i-1}$ then $R_i \geq A_i$.

We now use induction to show that $R_i \geq A = \min_i A_i$ for each $i = 1, \ldots H$. At the first link, if $C_0 < A_1$ then $R_1 = C_0$, but $C_0 \geq A$ and so $R_1 \geq A$. If $C_0 \geq A_1$ then $R_1 \geq A_1$, and so $R_1 \geq A$ because $A_1 \geq A$.

Suppose now that $R_k \geq A$, with $k < H$. If $R_k < A_{k+1}$, then $R_{k+1} = R_k$, but $R_k \geq A$ and so $R_{k+1} \geq A$. If $R_k \geq A_{k+1}$ then $R_{k+1} \geq A_{k+1}$, and so $R_{k+1} \geq A$ because $A_{k+1} \geq A$. Thus, $R \equiv R_H \geq A$. Q.E.D.

E. Summary of Packet-Train Dispersion

Let us summarize our findings for the packet-train bandwidth distribution $\mathcal{B}(N)$. As the packet-train length N increases, trains become more vulnerable to cross-traffic interference due to their larger duration. Thus, packet trains are not as appropriate as packet pairs to measure the capacity of a path. As N increases, the SCDR prevails in $\mathcal{B}(N)$, and the variability of $b(N)$ decreases. Additionally, if N is sufficiently large, the bandwidth measurement that corresponds to the mean packet-train dispersion, referred to as ADR, becomes independent of N. The ADR depends, in general, on the capacity and utilization of all links in the path, as opposed to the available bandwidth which depends only on the tight link, and it also depends on the routing of cross traffic relative to the measured path. Finally, the ADR is a lower bound for the path's capacity and an upper bound for the path's available bandwidth.

V. A Capacity-Estimation Methodology

We now present a capacity-estimation methodology that is based on the results of the previous two sections. This methodology has been implemented in a tool called *pathrate*.[6] The main results upon which *pathrate* is based on are given as follows.

- The optimal train length for detecting the capacity mode is $N = 2$, i.e., to use packet pairs.
- Using packet pairs with variable-sized packets makes the SCDR modes wider and weaker, facilitating the detection of a capacity mode.
- Using relatively larger (but still variable-sized) packets makes the PNCMs weaker, facilitating the detection of a capacity mode.
- The ADR can be measured with long packet trains (a few tens of packets). The ADR is a lower bound for the capacity of the path.

Pathrate requires the cooperation of both the sender and the receiver, i.e., it is a *double end-point methodology*. More flex-

[6] *Pathrate* is publicly available at http://www.pathrate.org.

ible approaches require access only at the sender, forcing the receiver to reply to each probing packet using ICMP, UDP-echo, or TCP-FIN packets. The drawback of those approaches is that the reverse path from the receiver to the sender, through which the replies are forwarded, can affect the bandwidth measurements. We prefer the double end-point methodology, even though it is less flexible, because it is more accurate.

Pathrate uses UDP for transferring probing packets. Additionally, *pathrate* establishes a TCP connection, referred to as *control channel*, between the sender and the receiver. We ignore any packet pairs or trains that encountered losses during the measurement process. As a simple form of congestion avoidance, *pathrate* aborts the measurement process when it detects a number of consecutive losses in the path. We ensure that the time interval between successive packet pairs or trains is larger than the round-trip time of the path, and not less than 500 msec. On the average, the probing traffic overhead during a run varies between 25–600 Kb/s, depending on the range of the bandwidth measurements.

Pathrate consists mainly of three execution phases. We describe the major tasks in each phase next.

Preliminary Measurements and Bin-Width Selection: Initially, *pathrate* detects the maximum train length $N_{\max}$ that the path can transfer without causing packet losses. The ADR estimation, during Phase II, uses trains of that length. Then, *pathrate* generates about 60 packet trains of gradually increasing length, from $N = 2$ to $N = \min(10, N_{\max})$ packets. These *preliminary measurements* are used in two ways.

First, from the preliminary measurements, *pathrate* calculates a reasonable *bandwidth resolution*, or *bin width*, ω. The bandwidth resolution ω is an important parameter for the detection of local modes in a distribution of bandwidth measurements (see the Appendix). ω is set to about 10% of the interquartile range of the preliminary measurements. Thus, a wider distribution of measurements leads to a larger bin width, in accordance with standard statistical techniques for density estimation [25]. The final capacity estimate of *pathrate* is a bandwidth range of width ω.

Second, the preliminary measurements terminate with a "Quick-Estimate," when there is very low variation. This happens at paths that are quite lightly loaded. Specifically, *pathrate* measures the coefficient of variation (CoV) of the preliminary measurements (ignoring some of the largest and smallest values). If CoV is less than a certain threshold, *pathrate* exits. The final capacity estimate $\tilde{C}$ in that case is calculated as the average of the preliminary measurements, after removing the 10% smallest and largest values.

Phase I: Packet-Pair Probing: In Phase I, *pathrate* uses a large number of packet pairs to uncover the local modes of the bandwidth distribution $\mathcal{B}$. We expect one of these modes to be the CM. Phase I consists of $K_1 = 1000$ packet-pair measurements with variable-sized probing packets. L varies between $L_{\min}$ and $L_{\max}$ bytes. The minimum size $L_{\min}$ results from the host-related timestamping constraints discussed at the end of Section III-C and is always larger than 550 bytes. The maximum size $L_{\max}$ is set to the control channel's maximum segment size.

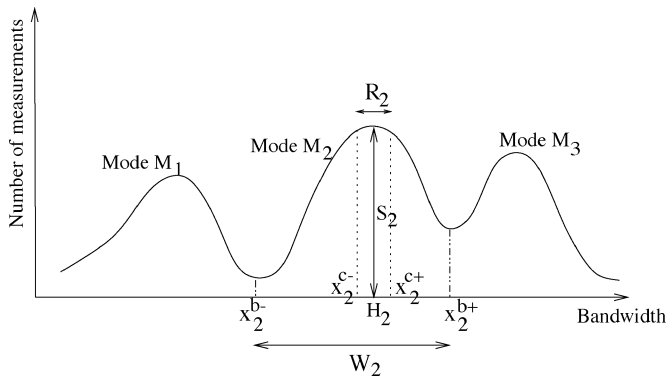

Fig. 12. Characteristics of a local mode.

At the end of Phase I, *pathrate* estimates the local modes in the distribution of packet-pair measurements. The Appendix describes the statistical procedure that we use for the identification of local modes. For each mode M_i, this procedure returns the central bin R_i, the number of measurements S_i in R_i, the range of the mode W_i, and the number of measurements B_i in W_i (see Fig. 12). The average bandwidth in the central bin R_i is denoted by H_i. The sequence of all local modes in the Phase-I measurements is $\mathcal{M} = \{M_1, M_2, \ldots M_\Pi\}$, with the modes ordered so that $H_i < H_{i+1}$.

We expect that one of these local modes, say M_k, is the CM, i.e., $H_k \approx C$. Modes M_i with $i > k$ are PNCMs, while modes M_i with $i < k$ reside in the SCDR. The challenge then is to select the right local mode M_k. We do so using an ADR estimate from Phase II.

Phase II: ADR Estimation and CM Selection: In Phase II, *pathrate* estimates the ADR from a number of long packet-train measurements. Specifically, Phase II consists of $K_2 = 500$ packet trains with length $N_{\max}$ and with maximum sized packets $(L = L_{max})$. Because $N_{\max}$ is several tens of packets, the resulting train bandwidth distribution $\mathcal{B}(N)$ has limited variability and it is typically unimodal. Using the same statistical procedure as in Phase I, *pathrate* estimates the ADR R as the global mode of this distribution.[7] Given that $R \leq C$, *pathrate* ignores all Phase-I modes M_i with $H_i < R$, because they probably are SCDR modes.

If there are more than one Phase-I modes that are larger than R, the next challenge is to select a mode that is most likely the CM. The insight that we use here is that CM is typically a relatively narrow and strong Phase-I mode. The reason is that Phase I uses variable-sized probing packets, widening the SCDR modes, and relatively large probing packets, weakening the PNCMs. CM, on the other hand, is not affected by either of these two techniques, and so it should be a relatively strong and narrow local mode.

To evaluate the strength and narrowness of Phase-I modes, we compute the following figure of merit F_i for each Phase-I mode:

$$F_i = S_i \Psi_i. \tag{19}$$

[7]We use the mode, rather than the mean, as the latter may be affected by outliers or any remaining local modes in $\mathcal{B}(N)$.

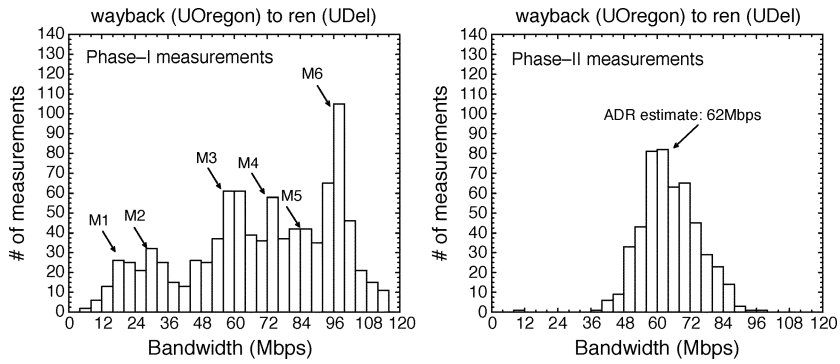

Fig. 13. Histograms of Phase-I and Phase-II measurements in an Internet path.

S_i is the number of Phase-I measurements in the central bin of mode M_i, and it estimates the strength of mode M_i. Ψ_i is the kurtosis of mode M_i, and it quantifies how narrow that mode is. The larger Ψ_i is the more "leptokurtic" mode M_i becomes, meaning that the measurements of M_i are more clustered around the center of that mode.

The final capacity estimate $\hat{C}$ is the Phase-I mode M_k that is larger (or equal) than R, and that has the maximum figure of merit F, i.e.,

$$k = \arg \max_{i=1...\Pi} F_i : \text{ such that } H_i \geq R. \tag{20}$$

Example of Operation: To illustrate the operation of *pathrate*, Fig. 13 shows the histograms of bandwidth measurements in Phase I and Phase II for a path that connects the University of Oregon to the University of Delaware. The local modes in Phase I are shown with arrows, while the ADR estimate in Phase II is about 62 Mb/s. Note that these two histograms are made with static bin-partitioning, and so they do not show the actual local modes that *pathrate* would estimate using the algorithm of the Appendix. The three Phase-I modes that are larger than the ADR are centered around 74, 83, and 98 Mb/s. The 98-Mb/s mode has the largest figrure of merit, and so it is the final *pathrate* estimate.

Accuracy of Pathrate: The first version of *pathrate* was released in the spring of 2000. Since then, the tool has gone through several revisions, and we have used it to measure hundreds of Internet paths. Here, we summarize our experiences with the accuracy of *pathrate*.

Overall, *pathrate* is quite accurate when the path meets the following two conditions: the capacity is not too high (typically below 500 Mb/s), and the path is not heavily loaded. Specifically, the tool has been consistently successful in measuring noncongested paths limited by Ethernet and Fast Ethernet segments, T1, T3, and OC-3 links, as well as slower links, such as dial-up, ADSL, and cable modems.

In high bandwidth paths, where the narrow link can be OC-12 (640 Mb/s) or Gigabit Ethernet (1000 Mb/s), *pathrate* can be less accurate, mostly due to the dispersion measurement noise at the receiver. The fact that *pathrate* uses application-layer timestamps can significantly deteriorate the accuracy of timing measurements when the latter are of the order of a few tens of microseconds. The use of a real-time operating system, as well as a high-resolution clock and timestamping facility, can improve the capacity estimation accuracy in high bandwidth paths [18]. The presence of interrupt coalescence at the receiver's network interface is another source of measurement error, but it can be dealt with as described in [22].

The accuracy of *pathrate* also drops in heavily loaded or congested paths. In that case, there may be no CM in the packet-pair bandwidth distribution, as almost all packet pairs encounter additional dispersion due to cross traffic. Typically, the accuracy of *pathrate* deteriorates when the utilization of the narrow link is more than 70%–80%. The use of smaller probing packets may be able to decrease the SCDR modes in that case, but it would also increase the intensity of PNCMs. Additionally, such an adaptive probing size selection would require some *a priori* knowledge of the load conditions in the measured path.

In terms of the estimation methodology that *pathrate* is based on, we found that most errors could be avoided with a different bin width ω. It is unfortunate that relatively small variations in this parameter often cause significant changes in the set of estimated local modes. The bin-width selection heuristic that we use, namely that ω is a fraction of the interquartile range of a certain number of measurements, is documented in the statistical literature (referred to as the "Freedman-Diaconis rule") [25], but nevertheless it is still a heuristic. A more robust (or adaptive) bin-width selection algorithm, or a fundamentally different technique to detect local modes in the packet-pair bandwidth distribution, are important problems for further investigation.

VI. Conclusion

Internet routers do not provide explicit feedback to end-hosts regarding the capacity or load of network links. Packet pairs and trains are useful probing mechanisms with which end-hosts can infer different bandwidth characteristics of a network path. This paper examined such packet-dispersion techniques, producing some negative and some positive results. A negative result is that it is difficult to measure the capacity of a path with just a few packet pairs. Another negative result is that packet trains do not measure the available bandwidth of a path, but a different metric (ADR) that is larger (or equal) than the available bandwidth. On the positive side, we showed that it is possible to estimate the capacity of a path with packet-dispersion techniques, especially if the path is not heavily loaded. To do so, it is im-

portant to understand the dispersion techniques not only in the statistical sense, but mostly in terms of the queueing effects that shape the distribution of bandwidth measurements. The main contribution of this study was to develop such an understanding, explaining the effect of various factors, including network load, cross-traffic packet-size variability, probing packet size, train length, and cross-traffic routing.

Appendix
Detection of Local Modes in a Multimodal Distribution

In general, the underlying distribution that generates a set of packet-pair bandwidth measurements can be multimodal. The local modes are the local maxima of the corresponding pdf. Since we only have a finite set of measurements, however, we can only approximate the pdf. In this Appendix, we describe a numerical algorithm that estimates the local modes of a distribution from a set of measurements.

Suppose that we have a set of K bandwidth measurements. We order the measurements in increasing sequence $x_1 \leq x_2 \leq \ldots \leq x_K$. The *rank* of a measurement is $r(x_m) = m$; ties for equal measurements are broken in an arbitrary manner. An important input parameter in the mode-detection algorithm is the *resolution* or *bin width* ω. We choose ω as a fraction (typically 10%) of the interquartile range of the measurements.

The following algorithm estimates iteratively a sequence of local modes $\{M_1, M_2, \ldots, M_\Pi\}$, with each iteration resulting in an additional mode. For each mode M_l, the following *mode characteristics* are reported (see Fig. 12):

- The range of the central bin $R_l = [x_l^{c-}, x_l^{c+}]$. The width of this range is no larger than ω.
- The number of measurements in R_l, $S_l = r(x_l^{c+}) - r(x_l^{c-}) + 1$.
- The range of the mode $W_l = [x_l^{b-}, x_l^{b+}]$. The mode includes all measurements distributed around the central bin R_l (described next).
- The number of measurements in W_l, $B_l = r(x_l^{b+}) - r(x_l^{b-}) + 1$.

The values x_l^{c-}, x_l^{c+}, x_l^{b-}, and x_l^{b+} for each local mode M_l are determined as follows.

Initially, all K measurements are unmarked and $l = 1$.

1) First, estimate the range R_l and the number of measurements S_l in the central bin of mode M_l. The defining property of the central bin is that it includes *the maximum number of consecutive and unmarked measurements in a range of width at most* ω. More formally, S_l is given by (21). x_l^{c-} and x_l^{c+} are the values of x_i and x_j, respectively, that determine S_l in the previous equation.
2) Next, find the right extent of mode M_l, i.e., estimate x_l^{b+}. This part of the algorithm is iterative. In each step, determine *the window at the right of the currently rightmost bin of* M_l, *overlapping with that bin, that includes the maximum number of measurements in a range of width at most* ω. If this window has more measurements than the rightmost bin of M_l, it belongs to a different local mode. Otherwise, set that bin as the rightmost bin of M_l, and repeat this step. The rightmost bin at the start of the iteration is the central bin of M_l:

$$S_l = \max_{1\leq i\leq K} \{t = j - i + 1 \text{ with } j = \arg\max_{i\leq j\leq K} \{x_j : x_j \leq x_i + \omega\} \text{ where } x_i, \ldots, x_j : \text{unmarked}\} \quad (21)$$

$$\hat{s} = \max_{i+1\leq m\leq j} \{t = n - m + 1 \text{ with } n = \arg\max_{m\leq n\leq K} \{x_n : x_n \leq x_m + \omega\}\}. \quad (22)$$

   More formally, suppose that the rightmost bin at iteration step k is $[x_i, x_j]$, with $s = j - i + 1$ measurements. Then, find the bin $[x_m, x_n]$ that resides at the right of $[x_i, x_j]$, overlapping with $[x_i, x_j]$, with the maximum number of measurements $\hat{s}$ in a range of width at most ω. $\hat{s}$ is given by (22). If $\hat{s} < s$, the bin $[x_m, x_n]$ is included in the mode M_l. In that case, set $m = i$, $n = j$, and $s = \hat{s}$, and repeat this iteration. Otherwise, $[x_i, x_j]$ marks the rightmost bin of mode M_l, and set $x_l^{b+} = x_j$.
3) Using the same approach as in the previous step, determine the leftmost bin of mode M_l, and compute x_l^{b-}.
4) Mark all measurements in mode M_l, from x_l^{b-} to x_l^{b+}. Then, set $l = l + 1$ and repeat the iteration from step 1). The algorithm terminates when all measurements are marked.

Note that, when searching for the next local mode, marked measurements will be skipped from the estimation of the central bin, which implies that the estimated local modes cannot have overlapping central bins. However, the range of adjacent modes may overlap.

The previous algorithm is based on *histograms of bandwidth measurements using a fixed bin width*. We have also experimented with two other statistical tools. The first is *histograms of dispersion measurements with a fixed bin width*. Such histograms were used, for instance, in [19]. Note that a fixed bin width at the dispersion domain corresponds to a variable bin width at the bandwidth domain, because dispersion is inversely proportional to bandwidth. So, the corresponding histogram at the bandwidth domain is *adaptive*, with an increasing bin width for larger bandwidth measurements. As one would expect, such adaptive histograms remove some SCDR modes (lower bandwidth values) by spreading those measurements into many bins, but they also intensify any PNCM modes (higher bandwidth values) by aggregating their measurements into wider bins. Overall, we found that adaptive histograms of this type do not lead to a consistent or significant improvement in the accuracy of *pathrate*.

Second, we experimented with a Kernel Density Estimator (KDE), using a fixed smoothing parameter at the bandwidth domain [25]. A KDE avoids the "origin-selection" problem of classical histograms. This technique has been used in [12], for instance. Note that, even though we do not need to arbitrarily select an origin for the estimated density, we still need to choose the equivalent of a bin width, referred to as the *smoothing parameter* of the KDE. In our experiments, we found that a KDE

does not help to identify the CM among the local modes of the packet-pair distribution, in the sense that a weak local mode in a histogram remains a weak local mode with a KDE, and a strong local mode in a histogram remains a strong local mode with a KDE.

ACKNOWLEDGMENT

The authors would like to thank R. Prasad, M. Jain, H. Jiang, W. Zhu, A. Dhamdhere, K. Claffy, and M. Murray, as well as the anonymous referees and the Transactions Editor for their many constructive comments.

REFERENCES

[1] I. C. Bolot, "Characterizing end-to-end packet delay and loss in the internet," in *Proc. ACM SIGCOMM*, Sept. 1993, pp. 289–298.
[2] L. S. Brakmo and L. L. Peterson, "TCP Vegas: end to end congestion avoidance on a global internet," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1465–1480, Oct. 1995.
[3] R. L. Carter and M. E. Crovella, "Measuring bottle-neck link speed in packet-switched networks," *Performance Eval.*, vol. 27/28, pp. 297–318, Oct. 1996.
[4] C. Dovrolis, P. Ramanathan, and D. Moore, "What do packet dispersion techniques measure?," in *Proc. IEEE INFOCOM*, Apr. 2001, pp. 905–914.
[5] A. B. Downey, "Using Pathchar to estimate internet link characteristics," in *Proc. ACM SICCOMM*, Sept. 1999, pp. 222–223.
[6] K. Harfoush, A. Bestavros, and J. Byers, "Measuring bottleneck bandwidth of targeted path segments," in *Proc. IEEE INFOCOM*, Mar. 2003, pp. 2079–2089.
[7] J. C. Hoe, "Improving the start-up behavior of a congestion control scheme for TCP," in *Proc. ACM SIGCOMM*, Aug. 1996, pp. 270–280.
[8] V. Jacobson, "Congestion avoidance and control," in *Proc. ACM SIGCOMM*, Sept. 1988, pp. 314–329.
[9] ——, (1997, Apr.) Pathchar: A Tool to Infer Characteristics of Internet Paths. [Online]ftp://ftp.ee.lbl.gov/pathchar/
[10] M. Jain and C. Dovrolis, "End-to-end available bandwidth: measurement methodology, dynamics, and relation with TCP throughput," *IEEE/ACM Trans. Networking*, vol. 11, pp. 537–549, Aug. 2003.
[11] S. Keshav, "A control-theoretic approach to flow control," in *Proc. ACM SIGCOMM*, Sept. 1991, pp. 3–15.
[12] K. Lai and M. Baker, "Measuring bandwidth," in *Proc. IEEE INFOCOM*, Apr. 1999, pp. 235–245.
[13] ——, "Measuring link bandwidths using a deterministic model of packet delay," in *Proc. ACM SIGCOMM*, Sept. 2000, pp. 283–294.
[14] S. McCreary and K. C. Claffy, "Trends in wide area IP traffic patterns: a view from AMES internet exchange," in *Proc. ITC Specialist Seminar IP Traffic Modeling, Measurement and Management*, Sept. 2000.
[15] B. Melander, M. Bjorkman, and P. Gunningberg, "A new end-to-end probing and analysis method for estimating bandwidth bottlenecks," in *Proc. IEEE Global Internet Symp.*, 2000.
[16] T. Oetiker. MRTG: Multi Router Traffic Grapher. [Online]. Available: http://eestaff.ethz.ch/~oetiker/webtools/mrtg/mrtg.html
[17] A. Pasztor and D. Veitch, "Active probing using packet quartets," in *Proc. Internet Measurement Workshop (IMW)*, Nov. 2002, pp. 293–305.
[18] ——, "PC based precision timing without GPS," in *Proc. ACM SIGMETRICS*, May 2002, pp. 204–213.
[19] ——, "The packet size dependence of packet pair like methods," in *Proc. IEEE/IFIP Int. Workshop Quality of Service (IWQoS)*, May 2002, pp. 204–213.
[20] V. Paxson, "Measurements and analysis of end-to-end Internet dynamics," Ph.D. dissertation, Univ. of California, Berkeley, Apr. 1997.
[21] R. S. Prasad, C. Dovrolis, and B. A. Mah, "The effect of layer-2 store-and-forward devices on per-hop capacity estimation," in *Proc. IEEE INFOCOM*, Mar. 2003, pp. 2090–2100.
[22] R. S. Prasad, M. Jai, and C. Dovrolis, "Effects of interrupt coalescence on network measurements," in *Proc. Passive and Active Measurtments (PAM) Workshop*, Apr. 2004.
[23] R. S. Prasad, M. Murray, C. Dovrolis, and K. Claffy, "Bandwidth estimation: metrics, measurement techniques, and tools," *IEEE Network*, vol. 17, pp. 27–35, Nov. 2003.
[24] V. Ribeiro, R. Riedi, R. Baraniuk, J. Navratil, and L. Cottrell, "PathChirp: efficient available bandwidth estimation for network paths," in *Proc. Passive and Active Measurements (PAM) Workshop*, Apr. 2003.
[25] D. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization*. Englewood Cliffs, NJ: Prentice-Hall, 1992.
[26] K. Thompson, G. J. Miller, and R. Wilder, "Wide-area internet traffic patterns and characteristics," *IEEE Network*, pp. 10–23, Nov. 1997.

Constantinos Dovrolis (M'93) received the Computer Engineering degree from the Technical University of Crete, Crete, Greece, in 1995, the M.S. degree from the University of Rochester, Rochester, NY, in 1996, and the Ph.D. degree from the University of Wisconsin-Madison in 2000.

He is an Assistant Professor with the College of Computing, Georgia Institute of Technology, Atlanta. His research interests include methodologies and applications of network measurements, bandwidth estimation algorithms and tools, overlay networks, service differentiation, and router architectures.

Prof. Dovrolis is a member of the Association for Computing Machinery.

Parameswaran Ramanathan received the B.Tech. degree from the Indian Institute of Technology, Bombay, in 1984, and the M.S.E. and Ph.D. degrees from the University of Michigan, Ann Arbor, in 1986 and 1989, respectively.

Since 1989, he has been a faculty member with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, where is presently a Full Professor. He leads research projects in the areas of sensor networks and next-generation cellular technology. His research interests include wireless and wireline networking, real-time systems, fault-tolerant computing, and distributed systems.

Dr. Ramanathan is presently an Associate Editor for the IEEE TRANSACTIONS ON MOBILE COMPUTING and the *AdHoc Networks Journal*. He served as an Associate Editor for IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED COMPUTING from 1996 to 1999.

David Moore (M'93) is currently working toward the Ph.D. degree at the University of California, San Diego.

He is a Principal Investigator and Assistant Director of the Cooperative Association for Internet Data Analysis (CAIDA). His research interests are high-speed network monitoring, denial-of-service attacks and infrastructure security, and Internet traffic characterization. In addition to network analysis, he has led several tool-development projects ranging from data collection to geographic mapping.

Mr. Moore is a member of the Association for Computing Machinery.