# Inferring Link Loss Using Striped Unicast Probes

N.G. Duffield† F. Lo Presti†,§ V. Paxson‡ D. Towsley§

†AT&T Labs–Research
180 Park Avenue
Florham Park, NJ 07932, USA
{duffield, lopresti}@research.att.com

‡ACIRI
International Computer Science Institute
Berkeley, CA 94704, USA
vern@aciri.org

§Dept. of Computer Science
University of Massachusetts
Amherst, MA 01003, USA
towsley@cs.umass.edu

*Abstract*— **In this paper we explore the use of end-to-end unicast traffic as measurement probes to infer link-level loss rates. We leverage off of earlier work that produced efficient estimates for link-level loss rates based on end-to-end multicast traffic measurements. We design experiments based on the notion of transmitting stripes of packets (with no delay between transmission of successive packets within a stripe) to two or more receivers. The purpose of these stripes is to ensure that the correlation in receiver observations matches as closely as possible what would have been observed if the stripe had been replaced by a notional multicast probe that followed the same paths to the receivers. Measurements provide good evidence that a packet pair to distinct receivers introduces considerable correlation which can be further increased by simply considering longer stripes. We then use simulation to explore how well these stripes translate into accurate link-level loss estimates. We observe good accuracy with packet pairs, with a typical error of about 1%, which significantly decreases as stripe length is increased to 4 packets.**

## I. Introduction

### A. Motivation

As the Internet grows in size and diversity, its internal performance becomes ever more difficult to measure. Any one organization has administrative access to only a small fraction of the network's internal nodes, whereas commercial factors often prevent organizations from sharing internal performance data.

One promising technique that avoids these problems, *Multicast Inference of Network Characteristics* (MINC), uses end-to-end multicast measurements to infer link-level loss rates and delay statistics by exploiting the inherent (and well characterized) correlation in performance observed by multicast receivers. These measurements do not rely on administrative access to internal nodes since the inference can be calculated using only information recorded at the end hosts.

The key intuition for inferring packet loss is that the arrival of a packet at a given internal node can be directly inferred from the packet's arrival at one or more receivers reached from the source by paths through that node; if it makes it to the receivers, it must have made it to the node. Conditioning on arrival at a descendent, we can determine the probability of successful transmission to and beyond the given node. Efficient inferencing algorithms are given in [2] for loss, [15] for delay distributions, [7] for delay variances, and [3] for inferring the logical multicast tree topology itself.

Although significant advances have been made in the use of multicast measurements for inferring internal network behavior, it suffers from two serious deficiencies. First, there remain significant portions of the Internet that do not support network-level multicast. Second, the internal performance observed by multicast packets often differs significantly from that observed by unicast packets. This is especially serious given that unicast traffic constitutes far and away the largest portion of the traffic on the Internet. Thus there is a need for techniques based on end-to-end unicast measurements. This poses a significant challenge because unicast measurements do not exhibit the well-behaved correlation exhibited by multicast. Thus, the challenge addressed in this paper is that of developing unicast-based measurement techniques that create sufficient correlation to yield fruitful inference.

### B. Contribution

In this paper we adapt the multicast inference techniques proposed in [2] to perform inference of internal network characteristics from unicast end-to-end measurements. The data for the inference comprises measured end-to-end loss of unicast probes sent from a source to a number of destinations. This is used to infer the loss and delay characteristics of each logical link of the source tree joining the source to the destinations, i.e., of the composite paths between its branch points.

The idea is to construct composite probes of unicast packets whose collective statistical properties closely resemble those of a multicast packet. We shall speak of **striping** a group of unicast packets across a set of destinations. This entails dispatching the packets back-to-back

This work was supported in part by DARPA and the AFL under agreement F30602-98-2-0238

from a source, each packet potentially having a different destination address. Our premise is that when the duration of network congestion events exceeds the temporal width of the stripe, packets should have very similar experience of the network upon traversing common portions of the paths to their destinations. If the experiences were identical, the packets from a stripe that attempt to traverse a given link would either all be lost, or encounter identical delay. Hence the packet loss and delays on a given link would be perfectly correlated within a stripe; the composite probe would have the same statistical properties as a notional multicast packet that followed the same source tree. In this case the methods of [2], [7], [15] could be applied immediately to infer the per link loss and delay statistics of the logical source tree.

However, correlations within stripes may be less than perfect in practice. This is because congestion events may not affect packets uniformly, subjecting stripes to dispersion as they travel through a network. Some mechanisms by which this can happen are the following. Packet loss will not be uniform during loss events that are narrower than the stripe, or those that start or stop while the stripe is in progress. Furthermore, delays will vary due to interleaving of background traffic, e.g., when moving from a low to a high capacity link. Although such effects should be small for sufficiently narrow stripes, they will be cumulative. Packet-dropping on the basis of Random Early Detection (RED) [9] is another mechanism by which packet loss may become decorrelated. It remains to be seen whether this mechanism will be widely deployed in communications networks. On the other hand, the use of RED to merely mark packets will not break correlations.

This motivates four strands of work in this paper:
*(i)* determining the magnitude of imperfect correlations through experiments on real networks;
*(ii)* calculating their likely impact on the accuracy of inference methods that assume perfect correlations;
*(iii)* adopting measurement procedures that reduce the impact of imperfect correlations;
*(iv)* verifying the accuracy of the approach in simulations.

We extend the packet loss model of [2] by incorporating an additional parameter for each link that describes the correlation of loss between different packets of the same stripe. This is done for binary stripes, i.e., those comprising two packets with different destination addresses. These additional parameters cannot themselves be determined by end-to-end measurements, at least not without additional assumptions relating them to each other, or to the existing loss rate parameters. These calculations show that the error in using the loss estimator from [2] is small provided that the conditional probability of *loss* of one packet in the stripe given *transmission* (i.e., non-loss) of the other, is small compared with the marginal loss rate in the stripe. This is a condition that we will verify, at least for end-to-end paths, through measurement.

By constructing appropriate stripes of composite probes and selecting subsets of these probes for inference, we are able to enhance correlations within data used for inference. This is possible when packet transmissions are correlated in the sense that a given packet in a stripe is more likely to be transmitted across a given link when other packets within the stripe are known to have been transmitted across the link. By conditioning on the measurable event that nearby packets have been transmitted end-to-end, we can raise the likelihood of transmission of a given packet to an intermediate node closer to one. By sending the stripe packets to diverse addresses, we can infer the properties of internal network paths from the measurements.

The rest of the paper is as follows. In Section II we formulate the stripe method, first for the binary tree of depth two, and then for general trees. We specify a family of different striping methods. We specify the required correlation assumption between packet transmissions within stripes, and show that it can be used to construct a hierarchy amongst the various striping methods; in particular we establish an order relation for the degree of correction each method gives to the bias caused by imperfect correlations.

We use two experimental approaches to evaluate the proposed method. In Section III we use end-to-end measurement on the National Internet Measurement Infrastructure (NIMI) [19] to gather data from a diverse set of Internet paths. We transmitted stripes between pairs of end-hosts and verified that their packet loss statistics were consistent with the correlation assumptions that underlie the method. (These stripes were different from those defined above, since all packets in the stripes were sent to the same destination; see Section III-A for discussion of this approach.) We also estimated the likely accuracy that would be obtained by stripe-based inference in the actual network.

We support this work in Section IV using network level simulation with `ns` [17]. By instrumenting the simulation we can trace the behavior of packets in the network interior. This allows us first to study the correlation properties of packets within stripes as they are transmitted across individual links in the network (rather than just the end-to-end properties), and second to compare the inferred link loss rates with actual link loss rates. For the most accurate choice of striping method we find the typical absolute error in loss rate inference to be below 1%. We conclude in Section V.

### C. Related Work

There exist several tools and methodologies for characterizing link-level behavior from end-to-end unicast measurements. One of the first methodologies focuses on identifying the bottleneck bandwidth on a unicast route. The key idea is that, in an uncongested network, two packets (packet pair) sent back-to-back will arrive at the receiver with a spacing that is inversely proportional to the lowest link bandwidth on the path. This was noted by Jacobson as leading to TCP's "self-clocking" behavior [10], and formally analyzed by Keshav [12]. Carter and Crovella then developed a tool to apply the technique [4], which has since been refined in [13], [18]. Although these methodologies focus on a metric other than loss rate, they are based on the same idea, namely to send packet pairs (or stripes) so as to introduce correlation in a controlled manner.

In [5], the authors use end-to-end measurements of packet pairs in a tree connecting a single sender to several receivers. Experiments consist of a number of packet pairs where the packets are sent to different receivers so that all pairs of receivers are covered. The metrics of interest are success probabilities of all links in the tree. As the second packet in a pair may not see the same loss behavior as the first over the common path, conditional success probabilities are introduced as unknown nuisance variables. Given an *a priori* distribution for these two sets of parameters, the authors then use a Bayesian network approach to determine *a posteriori* distributions and, from these, estimates of the link transmission probabilities. Preliminary results on the method reported in the paper show promise. Our approach differs from the approach in [5] in that we consider a more general form of striping scheme which results in significantly higher correlation. Thus we are able to continue to rely on the maximum likelihood estimates derived for the multicast case.

Last, `pathchar` [6], [11] triggers ICMP messages at successive routers on a unicast path in order derive link bandwidth, round trip link loss rate, and round trip link delay statistics. It accurately estimates link bandwidth provided that it is low. It has not been well validated in the case of losses and delays. Moreover, it requires considerable time to converge and loses accuracy with asymmetric round trip paths.

## II. INFERENCE METHODOLOGY

### A. Models for Trees, Stripes, and Packet Loss

We first develop the framework in which to describe the propagation of stripes of unicast packets through the network. We represent the underlying physical network as a graph $G_{\rm phys} = (V_{\rm phys}, L_{\rm phys})$ comprising the physical nodes $V_{\rm phys}$ (e.g. routers and switches) and the links $L_{\rm phys}$ between them. We consider a single source of probes $0 \in V_{\rm phys}$ and a set of receivers $R \subset V_{\rm phys}$. We assume that the set of paths from 0 to each $r \in R$ is stationary and form a tree $\mathcal{T}_{\rm phys}$ in $(V_{\rm phys}, L_{\rm phys})$; thus two such paths never intersect again once they have diverged. We form the logical source tree $\mathcal{T} = (V, L)$ whose vertices $V$ comprise 0, $R$ and the branch points of $\mathcal{T}_{\rm phys}$. The link set $L$ contains the link $(j, k)$ if one or more of the probe paths in $\mathcal{T}_{\rm phys}$ pass through $j$ then $k$ without encountering another element of $V$ in between. Where applicable, denote by $f(k) \in V$ the parent of $k \in V$. We write $j \succ k$ if $j$ is an ancestor of $k$ in $\mathcal{T}$.

We will use the notation $\langle r_1, \ldots, r_{d_0}\rangle$ to refer to a stripe comprising packets dispatched to destination nodes in order $r_1, \ldots, r_{d_0}$. We describe the progress of the stripe in $\mathcal{T}$ by the variables $X_k(d)$, taking the value 1 if packet $d$ reaches node $k$, and zero otherwise. Note $X_{r_d}(d) = 1$ iff packet $d$ reaches its destination node. (We do not label packets by their destination since we consider stripes with repeated destinations).

We will find it useful to have a notation describing composite events at sets of receivers. For $D \subset D_0 = \{1, \ldots, d_0\}$ define the binary variable

$$Z_D = \prod_{d \in D} X_{r_d}(d). \tag{1}$$

Thus $Z_D = 1$ if all packets in $D$ reach their destinations, and 0 otherwise. We will find it convenient to write $Z_{\{d_1,\ldots,\, d_m\}}$ as $Z_{d_1\ldots\, d_m}$.

We specify a loss model for the stripes. We assume that losses are independent between different stripes, and for packets of the same stripe on different links. For each $k \in V$ let $D(k) \subset D_0$ be the set of packets that successfully reach (and therefore transit across) $k$. For $D \subset D(k)$ let $\alpha_k(D)$ denote the probability that all packets in $D$ are transmitted to node $k$, conditioned upon having reached the parent node $f(k)$. We do not assume that the marginal probabilities $\alpha_k(d)$ are equal for all $d \in D(k)$. For disjoint subsets $D, D' \subset D(k)$ we write as $\beta_k(D|D')$ the conditional probability that packets in $D$ are successfully transmitted across link $k$, given that those in $D'$ are successfully transmitted, all packets having reached the parent node $f(k)$. This is expressed in terms of the probabilities $\alpha_k$ as

$$\beta_k(D|D') = \alpha_k(D \cup D')/\alpha_k(D'). \tag{2}$$

With perfect correlations the various $\beta_k$ would be 1. The multicast loss model of [2] is statistically equivalent to the special case $\beta_k(D|D') = 1$ and hence $\alpha_k(d)$ all equal some $\alpha_k$.

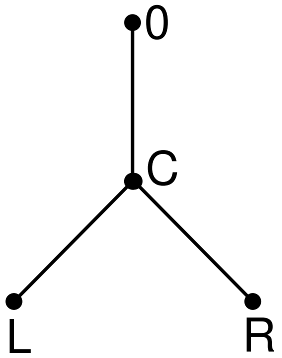


Fig. 1. TWO-LEAF TREE

For a given link and stripe width, we expect the structure of the probabilities $\alpha, \beta$ to depend on the times between successive packets. For example, if the packets are widely separated, then the marginal probabilities $\alpha_k(d)$ will be equal (or nearly so) while the conditional probabilities $\beta$ will be close to the marginal probabilities $\alpha$. Here, we concentrate on the other extreme with back-to-back packets in order to make $\beta$ close to 1. In this paper we focus on estimating transmission probabilities for the first packet in a stripe. We note however that marginal transmission probabilities can depend on the position of a packet within a stripe, particularly when the stripe width is not negligible compared with buffer sizes. However, our methods can be adapted to focus on other packets within the stripe. This could be useful if it is desired to infer transmission probabilities for packets in traffic bursts.

### B. Inference with Binary Stripes on the Two-Leaf Tree

We first investigate the performance of the inference algorithms from [2] under imperfect correlations. We start with the two-leaf tree shown in Figure 1, having leaf nodes L and R with common parent C whose own parent is the root 0. Consider the binary stripe $\langle \mathrm{L}, \mathrm{R} \rangle$. The link probabilities are related to the probabilities of leaf events as follows:

$$\frac{\mathsf{E}Z_1 \mathsf{E}Z_2}{\mathsf{E}Z_{12}} = \alpha_{\mathrm{C}}(1)/\beta_{\mathrm{C}}(1|2) \tag{3}$$
$$\frac{\mathsf{E}Z_{12}}{\mathsf{E}Z_2} = \alpha_{\mathrm{L}}(1)\beta_{\mathrm{C}}(1|2), \qquad \frac{\mathsf{E}Z_{12}}{\mathsf{E}Z_1} = \alpha_{\mathrm{R}}(2)\beta_{\mathrm{C}}(2|1),$$

where $Z_D$ is as defined in (1). This is because, e.g., $\mathsf{E}Z_{12} = \alpha_{\mathrm{C}}(12)\alpha_{\mathrm{L}}(1)\alpha_{\mathrm{R}}(2) = \alpha_{\mathrm{C}}(1)\beta_{\mathrm{C}}(2|1)\alpha_{\mathrm{L}}(1)\alpha_{\mathrm{R}}(2)$, with similar expressions for $\mathsf{E}Z_1$ and $\mathsf{E}Z_2$. With perfect correlations, $\beta_{\mathrm{C}} = 1$, and hence the $\alpha$ are uniform across the stripe and may be recovered directly from the leaf probabilities. These expressions can then be used to estimate the $\alpha$ from the leaf events $Z^{(i)}$ associated with multiple identical stripes $i = 1, 2, \dots n$. To form the estimates we first replace each expectation in (3) by the corresponding empirical mean, defined here in general:

$$\widetilde{Z}_D = n^{-1} \sum_{i=1}^{n} Z_D^{(i)}. \tag{4}$$

Taking $\beta_{\mathrm{C}} = 1$ then yields the estimates

$$\widehat{\alpha}_{\mathrm{C}} = \widetilde{Z}_1 \widetilde{Z}_2 / \widetilde{Z}_{12}, \ \widehat{\alpha}_{\mathrm{L}} = \widetilde{Z}_{12}/\widetilde{Z}_2, \ \widehat{\alpha}_{\mathrm{R}} = \widetilde{Z}_{12}/\widetilde{Z}_1. \tag{5}$$

This is effectively the estimator from [2] applied to the two-leaf tree.

With imperfect correlations, $\beta_{\mathrm{C}}$ cannot be recovered independently from the leaf expectations. The model is not identifiable; this was also observed in [5]. Since $\beta_{\mathrm{C}} \le 1$, estimation via (5) is biased, overestimating $\alpha_{\mathrm{C}}$ and underestimating $\alpha_{\mathrm{L}}$ and $\alpha_{\mathrm{R}}$.

### C. Enhancing Stripe Correlations

The uncertainty over the values of the $\beta$ undermines confidence in using (5) directly. We now propose a modified striping scheme scheme for which the effective value of the $\beta$ is closer to 1. To glimpse the idea behind this, observe that for the stripe $\langle \mathrm{L}, \mathrm{R} \rangle$ with perfect correlations, $\mathsf{E}Z_{12}/\mathsf{E}Z_2$ (defined as the conditional probability for the first packet of the stripe to reach L given that its second packet reaches C) is actually equal to the probability of transmission of a packet along the link $(\mathrm{C}, \mathrm{L})$, conditional upon reaching C. This is because packet 2 must have been present at C if present at R. With imperfect correlations, packet 1 may not have been also present at C, leading to underestimation of $\alpha_{\mathrm{L}}$. Our remedy for this is to use longer stripes, conditioning on an event at R which makes it more likely that packet 1 was present at C.

The simplest example of such a stripe is the **three-packet stripe** $\langle \mathrm{L}, \mathrm{R}, \mathrm{R} \rangle$. Provided that transmission of packets within the stripe is strongly correlated (in a sense we make precise below) we expect it to be more likely that packet 1 reaches C, upon reception of packets 2 and 3 at receiver L, rather than reception of packet 2 alone. We formalize the required notion of correlation in Definition 1 below.

Upon replacing the reception of packet 2 with the reception of packets 2 and 3, the analogs of the first and second relations in (3) are

$$\frac{\mathsf{E}Z_1 \mathsf{E}Z_{23}}{\mathsf{E}Z_{123}} = \frac{\alpha_{\mathrm{C}}(1)}{\beta_{\mathrm{C}}(1|23)}, \ \frac{\mathsf{E}Z_{123}}{\mathsf{E}Z_{23}} = \alpha_{\mathrm{L}}(1)\beta_{\mathrm{C}}(1|23). \tag{6}$$

The parameters $\alpha_{\mathrm{C}}$ and $\alpha_{\mathrm{L}}$ are estimated by $\widetilde{Z}_1 \widetilde{Z}_{23}/\widetilde{Z}_{123}$ $\widetilde{Z}_{123}/\widetilde{Z}_{23}$ respectively; $\alpha_{\mathrm{R}}$ can be estimated similarly using the complementary stripe $\langle \mathrm{R}, \mathrm{L}, \mathrm{L} \rangle$. Comparing with

(5) we observe that these estimates introduce less bias than those from two-packet stripes provided that $\beta_{\mathrm{C}}(1|2,3) > \beta_{\mathrm{C}}(2|1)$. This is the case provided that transmissions within a stripe satisfy the following correlation property.

*Definition 1:* We say that stripe transmission at a node $k$ is **coalescent** if for each stripe $\langle r_1,\ldots,r_d\rangle$ routed through $k$, and disjoint $D, D' \subset D(k)$,

$$\beta_k(D|D') \geq \beta_k(D|D'') \text{ for all } D'' \subset D'. \tag{7}$$

Coalescence is a correlation property. It states that a given set of packets $D$ is more likely to be transmitted on a link, the more other packets from the stripe have been transmitted. We will investigate the coalescence properties of real network traffic in Section III.

With coalescence, whenever we add packets to the conditioning event, the effect is to decrease the estimate of $\alpha_{\mathrm{C}}$ and to increase the estimate of $\alpha_{\mathrm{L}}$ or $\alpha_{\mathrm{R}}$. Thus, we can counteract the bias in the two-leaf stripe, evident from (3), by using wider stripes.

*Theorem 1:* Assume transmission is coalescent on the two-leaf tree and consider a stripe $\langle D(\mathrm{C})\rangle$ and two disjoint subsets $D, D'$ of $D(\mathrm{C})$ such that packets in $D$ have destination $\mathrm{L}$ and packets in $D'$ have destination $\mathrm{R}$. Then for any $D'' \subset D'$,

$$\frac{\mathsf{E}Z_{D\cup D'}}{\mathsf{E}Z_{D'}} \geq \frac{\mathsf{E}Z_{D\cup D''}}{\mathsf{E}Z_{D''}}. \tag{8}$$

The inequality (8) captures the effect that extending the stripe reduces the estimate of the transmission rate $\alpha_{\mathrm{C}}$ and so counteracts the bias due to $\beta_{\mathrm{C}} < 1$.

*Proof:* $\mathsf{E}Z_{D\cup D'} = \beta_{\mathrm{C}}(D|D')\alpha_{\mathrm{C}}(D')\,\alpha_{\mathrm{L}}(D)\alpha_{\mathrm{R}}(D')$ while $\mathsf{E}Z_{D'} = \alpha_{\mathrm{C}}(D')\alpha_{\mathrm{R}}(D')$. Hence $\mathsf{E}Z_{D\cup D'}/\mathsf{E}Z_{D'} = \beta_{\mathrm{C}}(D|D')\alpha_{\mathrm{L}}(D) \geq \beta_{\mathrm{C}}(D|D'')\alpha_{\mathrm{L}}(D) = \mathsf{E}Z_{D\cup D''}/\mathsf{E}Z_{D''}$. ■

*Example: the 4-packet stripe.* Theorem 1 suggests we can further reduce bias by lengthening the stripe length. Consider, for instance, the stripe $\langle \mathrm{L},\mathrm{R},\mathrm{R},\mathrm{R}\rangle$ and compare its estimation properties with those of its substripes $\langle \mathrm{L},\mathrm{R},\mathrm{R}\rangle$ and $\langle \mathrm{L},\mathrm{R}\rangle$. By Theorem 1 we have the following ordering between the functional on which estimates of $\alpha_{\mathrm{C}}$ are based in each case:

$$\frac{\mathsf{E}Z_1\mathsf{E}Z_{234}}{\mathsf{E}Z_{1234}} \leq \frac{\mathsf{E}Z_1\mathsf{E}Z_{23}}{\mathsf{E}Z_{123}} \leq \frac{\mathsf{E}Z_1\mathsf{E}Z_2}{\mathsf{E}Z_{12}}. \tag{9}$$

The estimators are obtained by replacing each $\mathsf{E}Z$ by the corresponding empirical mean $\widetilde{Z}$ from $n$ stripes. By the Law of Large Numbers, the same inequalities hold for the estimates with probability 1 as $n$ grows to infinity.

### D. Extension to General Trees

We describe estimators that extend the foregoing method to treat general logical source trees, i.e., trees in which the depth and branching ratio can be greater than 2. Consider first the case of a depth 2 tree with an arbitrary number of leaves. One approach is to stripe across all receivers and then to adapt the estimator from [2] for nodes with arbitrary numbers of offspring in order to estimate the link probabilities. A potential problem with this approach is that the statistical properties of stripes may not reflect those of general traffic if their width is not negligible compared with buffer sizes. For the same reason, variation of stripe width within a single set of measurements may introduce non-uniform bias into the link probability estimates, depending on the local branching ratio. Instead, here we focus on combining inference from fixed-width stripe measurements on embedded subtrees.

Consider an arbitrary tree with leaf set $R$. For each node $k$ let $R(k)$ denote the subset of leaves descended from $k$. Let $Q(k)$ denote the set of ordered pairs of nodes in $R(k)$ descended through different children of $k$. For each $(\mathrm{R}_1,\mathrm{R}_2) \in Q(k)$, consider the embedded two-leaf binary tree spanned by the nodes $0, k, \mathrm{R}_1, \mathrm{R}_2$. By combining estimates from measurements of stripes down each such tree, we shall estimate the characteristics of the common path from 0 to $k$.

Each stripe will follow the same pattern. We fix a template for a stripe of $d_0$ packets by partitioning $\{1,\ldots,d_0\}$ into two sets $D_1, D_2$. For each ordered pair $(\mathrm{R}_{i_1},\mathrm{R}_{i_2})$ of distinct receivers in $R(k)$ we form a stripe that sends packets in positions in $D_1$ to $\mathrm{R}_{i_1}$ and packets in positions in $D_2$ to $\mathrm{R}_{i_2}$. More formally, this is the stripe $S(i_1,i_2) = \langle r_1,\ldots,r_{d_0}\rangle$ where $r_d = \mathrm{R}_{i_k}$ when $d \in D_k$.

The relation between the leaf probabilities and the transmission probabilities on the composite path from 0 to $k$ are expressed through

$$\frac{\mathsf{E}Z_{D_1}\mathsf{E}Z_{D_2}}{\mathsf{E}Z_{D_1\cup D_2}} = A_k(D_1)/B_k(D_1|D_2). \tag{10}$$

where $A_k = \prod_{j\succeq k}\alpha_j$ and $B_k = \prod_{j\succeq k}\beta_j$. For each non-leaf and non-root node $k$, each pair $(i,j) \in Q(k)$, the measurements with $n$ stripes of type $S(i,j)$ thus gives rise to an estimate

$$\widehat{A}_k^{i,j} = \frac{\widetilde{Z}_{D_1}\widetilde{Z}_{D_2}}{\widetilde{Z}_{D_1\cup D_2}}. \tag{11}$$

In the experiments described in this paper we combine all possible estimates through their arithmetic mean

$$\widehat{A}_k = \#Q(k)^{-1}\sum_{(i,j)\in Q(k)}\widehat{A}_k^{i,j}. \tag{12}$$

For leaf nodes $k$ take $\widehat{A}_k$ as the measured transmission probability over all stripes of packets to $k$, and set $\widehat{A}_0 = 1$

by convention. The link probability estimates are then expressed as quotients

$$\widehat{\alpha}_k = \widehat{A}_k / \widehat{A}_{f(k)}, \quad k \neq 0. \tag{13}$$

### *E. Sampling and Statistical Issues*

Earlier in this section we proposed using wider stripes as a way of counteracting the inherent bias in using estimators that do not take explicit account of the imperfect correlations between stripe packets. We now make a number of further observations of the statistical implications of using the stripe approach.

First, increasing the stripe width while keeping the total number of packets sent constant increases the variance of the estimates. This is because the number of stripes sent is in inverse proportion to their width.

Second, network characteristics may not be uniform across a stripe e.g., if stripe width is comparable in size to that of a buffer. Here we focused on estimating transmission probabilities for the first packet; other templates could direct attention to other positions. We note that if marginal transmission rates are highly heterogeneous across different positions in a stripe, then the assumption of independent packet loss on different links may not hold. This is because its expected loss rate of a packet at a given node can depend on the occurrence of losses closer to the source of packets in earlier stripe positions. These cause the packet to advance its position in the stripe and consequently experience a different loss rate.

Third, there is a phenomenon during TCP slow start that can lead to every other or every third packet being lost. Once TCP increases its window enough to "fill the pipe," which corresponds to transmitting at the bottleneck rate, then the next set of acknowledgements effectively increases the sending rate by either a factor of two (if the receiver acknowledges every incoming packet) or a factor of 1.5 (if the receiver uses the common "ack every other" policy). If the bottleneck buffer is full at this point, then either every other or every third packet will be lost at the bottleneck due to the mismatch between the bottleneck rate and the higher sending rate. See Figure 2 of [8] for an illustration. Accordingly, there may be buffer-filling patterns present in the network that impart particular loss patterns on the elements of a stripe. The prevalence of the "slow start" pattern will depend on how often TCP connections in slow start dominate the consumption of buffer space at the bottleneck link.

Fourth, we have observed that imperfect correlations at a node bias inference for parent and child links in opposite directions. Hence bias is a second order effect spatially, depending not on the absolute loss correlation, but rather on the manner in which it changes from node to node in the network. In the special case of the probabilities $\alpha, \beta$ being uniform over all links, imperfect correlations actually leave the estimates (5) *unbiased* for internal links (i.e. all those except the leaf links and root link), though this special case seems highly unlikely in practice.

Fifth, the analysis of estimator variance for multicast inference carries over when $\beta \approx 1$. We refer the reader to [2] for details. Here we mention that in a regime for which all loss rates $\overline{\alpha}_k = 1 - \alpha_k$ are close to zero, the estimator $\widehat{\alpha}_k$ has variance which behaves as $n^{-1}\left(\overline{\alpha}_k + \|\overline{\alpha}\|^2\right)$, asymptotically for large numbers $n$ of probes. To leading order, this form is independent of topology.

## III. Network Experiments

The estimation techniques described in Section II rely on conditional probabilities of packet transmission within stripes being close to 1, and on the coalescence property in order to counteract the bias due to shortcomings with this assumption. In this section we investigate conformance of both of these assumptions to measurements of stripes transmitted across a number of end-to-end paths in the Internet. Although these experiments did not access the transmission properties of individual links (logistically very difficult to measure), they would be able to detect link-wise departures from the assumptions, since these would also be reflected in the properties of end-to-end paths over non-conformant links.

### *A. Measurement Infrastructure*

We conducted the experiments using the National Internet Measurement Infrastructure (NIMI) [19]. NIMI consists of a number of measurement platforms deployed across the Internet (primarily in the U.S.) that can be used to perform end-to-end measurements. We made the measurements using the `zing` utility, which sends UDP packets in selectable patterns, recording the time of transmission and reception. We extended `zing` to transmit unicast stripes to multiple destinations, minimizing the spacing between packets in a stripe by precomputing the packets to send (including their MD5 integrity checksum, the most computationally expensive part of constructing a `zing` packet) and then transmitting them with back-to-back system calls, resulting in inter-packet spacings of about $40\mu$sec.

A key point is that for our measurements we did *not* actually send packets to multiple destinations, because we had no way of calibrating true inference of internal loss characteristics, which would require measurement inside the network. Instead, the results we report are all for stripes sent to the *same* destination, with the goal being

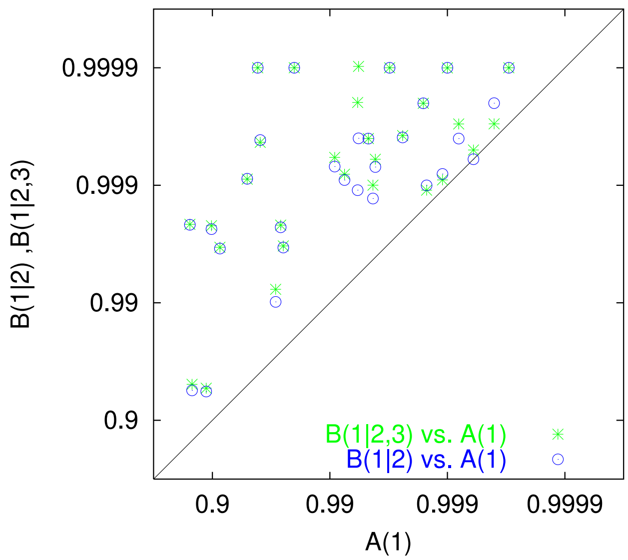


Fig. 2. SCATTER PLOT OF TRANSMISSION PROBABILITIES IN 28 NETWORK EXPERIMENTS. Conditional vs. marginal end-to-end transmission probabilities. Probabilities for 3-packet stripes mostly meet or exceed those for 2-packet stripes.

to assess the conditional loss probability and coalescence properties.

We gathered a total of 63 successful measurements between 35 NIMI sites, each measurement recording at both sender and receiver the transmission of either 100,000 flights of stripes of 3 packets, with separations exponentially distributed with a mean of 100 msec; 10,000 flights of stripes of 10 packets, separated by a mean of 300 msec (we also analyzed the first 3 packets in each stripe as another dataset of 3-packet stripes); or 20,000 flights of stripes of 3 packets, separated by a mean of 500 msec. All measurements were made at either 2PM EDT (a busy time) or 2AM EDT (a fairly unloaded time). There was no noticeable change in behavior as we varied the inter-stripe spacing from 100 msec to 500 msec.

Of the 63 traces, 7 exhibited no loss whatsoever, and consequently we had to eliminate them as they could not be used to study loss inference. Of the remaining 56, fully half (28) had conditional loss probabilities of 1, reflecting perfect loss correlation just as we would have if using multicast traffic instead of unicast. This finding is highly encouraging for the efficacy of unicast loss inference.

In the remainder of this section, we analyze the properties of the 28 traces that did not exhibit perfect correlation.

| | $\widehat{B}(1\|2,\ldots,w)/\widehat{B}(1\|2,\ldots,w-1)$ | | | | |
|---|---|---|---|---|---|
| | $w=2$ | $w=3$ | $w=4$ | $w=5$ | $w=6$ |
| min. | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| mean | 1.0189 | 1.0002 | 1.0000 | 1.0001 | 1.0001 |
| max. | 1.1812 | 1.0021 | 1.0003 | 1.0005 | 1.0003 |

TABLE I
COALESCENCE OF TRANSMISSION IN NETWORK EXPERIMENTS. RATIOS OF END-TO-END CONDITIONAL TRANSMISSION PROBABILITIES IN STRIPES OF WIDTH 2 TO 6. MINIMUM, MEAN AND MAXIMUM OF RATIOS OBSERVED IN 19 TRACES STRIPES OF WIDTH 10. MINIMUM RATIO 1 CONFORMS WITH COALESCENCE PROPERTY.

### B. *Transmission Probabilities*

*Marginal Probabilities.* The packet loss rate varied between zero and about 14% over the experiments. The marginal packet loss rates for different positions in the stripe displayed some heterogeneity. The heterogeneity was most pronounced at the start of the stripe, with the loss rate for the second packet in a stripe being typically 1.19 times greater than that of the first. Moving further along the stripe, loss rates differed between successive positions typically by up to a typical factor of 1.03.

*Conditional Probabilities.* We can estimate the error involved in the stripe method by comparing conditional and marginal transmission probabilities within the stripe. A scatter plot of the conditional vs. marginal probabilities for 2 and 3 packet stripes in 28 experiments is shown in Figure 2. Higher points represent smaller relative error; conversely for points near the line the error is of the same order of magnitude as the marginal probability to be estimated. For both 2 and 3 packet stripes, the end-to-end conditional transmission probabilities $\widehat{B}$ are noticeably larger than the marginal transmission probabilities $\widehat{A}$, with those for the 3 packet stripe being at least as large as those for the 2 packet stripes in almost all cases. A conditional probability of 1 would signify perfect correlations. We can characterize this error arising from $\widehat{B} < 1$ through the ratio $(1-\widehat{B})/(1-\widehat{A})$ when $\widehat{A} \neq 1$. This represents the proportion of the reported loss rate which is typically in error due to imperfect correlations. For 2-packet stripes, the median value of this ratio was 0.12. (So, for example, an estimated loss rate of 1% would be in error by about 0.12%). The median ratio fell to to 0.09 for 3 packet stripes.

*Coalescence* We calculated end-to-end conditional transmission probabilities $\widehat{B}(1|2,3,\ldots w)$ for stripes of width $w$ between 1 and 6. (When $w=1$ this just denotes the marginal probability $\widehat{A}(1)$). A necessary condition for coalescence is that the ratios $\widehat{B}(1|2,\ldots,w)/\widehat{B}(1|2,\ldots,w-$

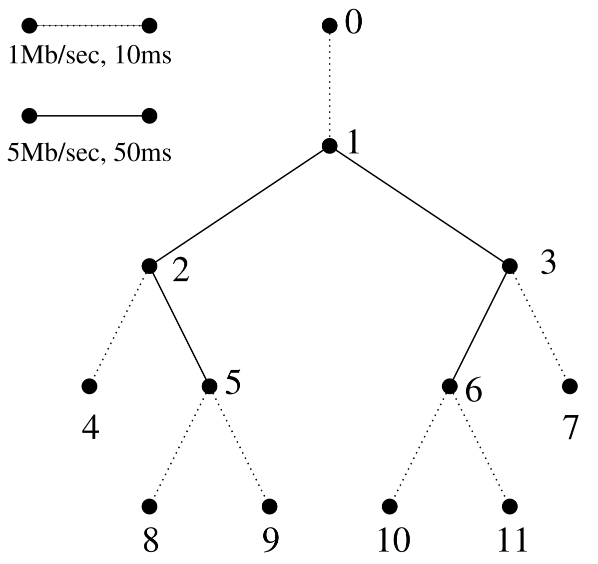


Fig. 3. SIMULATION TOPOLOGY

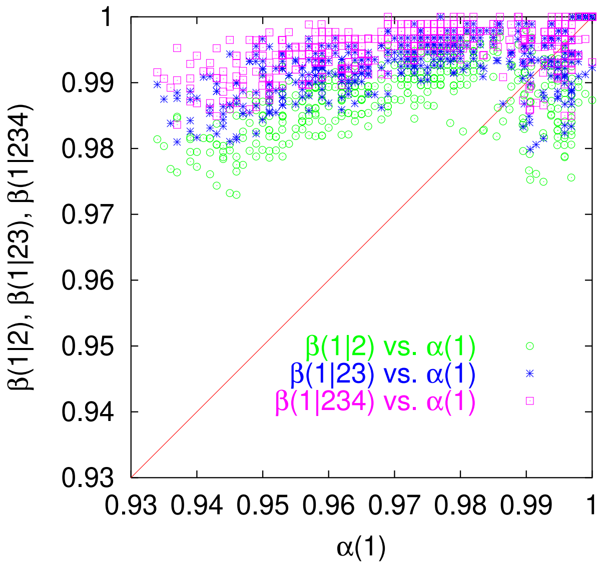


Fig. 4. CONDITIONAL TRANSMISSION PROBABILITIES IN SIMULATIONS. Scatter-plot of conditional vs. marginal link transmission probabilities for 2, 3 and 4 packet stripes. Conditional probabilities increase with stripe width.

1) be $\geq 1$. We determined the ratios over 19 experiments with stripes of width 10. In only two instances were the ratios less than 1, and in these cases by a magnitude of only about $10^{-6}$. This is a far smaller magnitude than that by which the ratio typically exceeds 1, as is seen from the statistics displayed in Table I: the minimum, mean, and maximum for each $w$ over the 19 experiments. The ratios are largest for $w = 2$, falling off close to 1 as $w$ increases beyond 3. This suggests that the additional bias correction obtained by increasing stripe width is almost negligible for stripes wider than 3 packets, at least under the network conditions and the range of loss probabilities exhibited in these traces.

### C. Interpretation

The network experiments are encouraging for unicast-based inference. First, in half of the traces the stripes exhibited perfect correlations. If this property were reproduced in stripes to multiple destinations, their statistical properties would be identical to that of multicast traffic for the purposes of link loss inference. Second, in traces with imperfect correlations, the conditional transmission probabilities within the stripe were considerably higher than the marginal probabilities, slightly more for the 3 packet stripe than the 2 packet stripe. This indicates that the bias due to ignoring the imperfection in correlations is relatively small. Third, traces exhibited coalescence for the stripe widths considered, indicating that the bias can be compensated for by using wider stripes, although the incremental benefit grew smaller for larger stripe widths. These factors lead us to expect that striped unicast probing will be quite effective for loss inference under real network conditions.

## IV. SIMULATION RESULTS

### A. Methodology

The experiments of Section III give us confidence that the statistical properties of stripe transmission make stripes suitable as probes for inference. However, the experiments do not enable us to corroborate the accuracy of the estimators for real network traffic. Instead, we employ simulation to get a sense of how accurate the estimators might be in practice.

We used the `ns` simulation environment [17]; this enables the representation of transport-protocol detail of packet transmissions, with packet loss due to buffer overflows at nodes as stripes compete with background traffic. The simulations reported in this paper used the topology of Figure 3. The different link speeds and delays are intended to characterize low speed/low delay links at a network edge connected by high speed/high delay links in the network interior. The goal is to study the methodology in a simplified environment to look for major problems, not to make a definitive assessment of the methodology.

Background traffic comprised a mixture of sessions over TCP and exponential on-off sources. There were on average 11 sessions per link direction. The buffer on each link accommodated 20 packets. Measurement probes comprised stripes with a $1\mu$sec interpacket time. Stripes were generated periodically with an inter-stripe of 16 msec. The tree was covered by cycling through thirty stripes $S(i, j)$

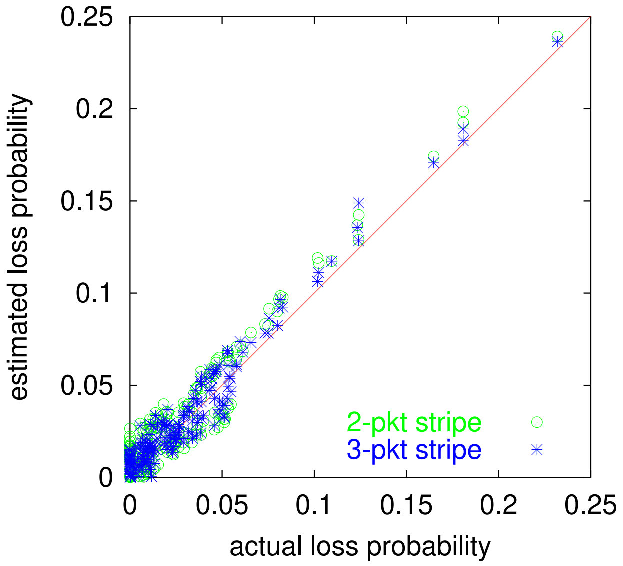


Fig. 5. Inferred vs. actual link loss rates in simulations. 3 packet and 2 packet substripes. Scatter plot for 100 experiments.

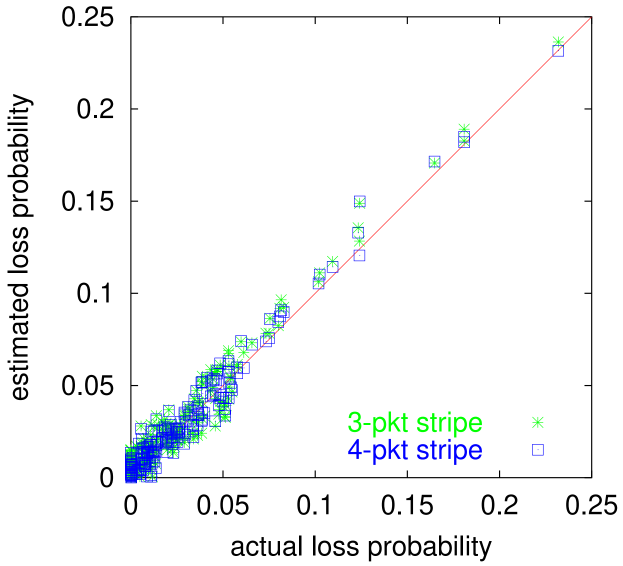


Fig. 6. Inferred vs. actual link loss rates in simulations. 4 packet stripes and 3 packet substripes. Scatter plot for 100 experiments.

over pairs of distinct receivers $i, j$. During an experiment, each stripe was transmitted 1,000 times. We conducted a set of 100 experiments using 4 packet stripes. To compare the estimator performance under the different stripe lengths we considered the 2 and 3-packet substripes obtained using the first two and three packets in each stripe. In order to evaluate the method, the inferred loss rates were compared with internal link loss rate as determined by instrumentation of the simulation. Link loss rates were computed considering only the first probe in the stripe.

### *B. Conditional and Marginal Transmission Probabilities*

We first examine the statistical properties of the underlying link loss processes. Figure 4 is a scatter plot of conditional vs. marginal transmission probabilities for 2, 3 and 4 packet stripes. Observe that conditional probabilities increase with stripe width. We summarize the likely relative errors in each case though the statistics of the ratio $(1-\widehat{\beta})/(1-\widehat{\alpha})$ of conditional to marginal loss probabilities. For 2 packet stripes the median ratio was 0.32 (i.e., a relative error of 32%). The ratio fell to 0.20 for 3 packet stripes, and further to 0.12 for 4-packet stripes.

These errors are somewhat greater than those observed for end-to-end transmission in the network experiments. We believe this may be associated with a greater heterogeneity in marginal transmission rates that we observed in the simulations; loss rates grew by about 30% between successive positions for the first 4 packets of a stripe. Recall from Section III-B that in the network experiments, the largest such ratio was 19%, and typical ratios were 3%.

| | stripe width | | |
|---|---|---|---|
| | 2 | 3 | 4 |
| mean | 0.0099 | 0.0075 | 0.0063 |
| s.d. | 0.0064 | 0.0057 | 0.0052 |

TABLE II

ESTIMATION ERROR IN SIMULATIONS AS FUNCTION OF STRIPE WIDTH. MEAN AND STANDARD DEVIATION OF ABSOLUTE DIFFERENCE BETWEEN INFERRED AND ACTUAL LOSS RATES. ERRORS ARE MINIMIZED FOR 4-PACKET STRIPES.

The stronger growth in loss ratios along the stripe in the simulations may be due to the larger size of the stripe relative to buffer size (20 packets) as compared with that in real networks.

### *C. Accuracy of Inference*

Finally, we compare inferred and actual link loss rates in the simulations. We display scatter plots of inferred vs. actual loss for 2 and 3 packet stripes in Figure 5, and 3 and 4 packet stripes in Figure 6. The same number of stripes was used in each case. From the figures we observe that accuracy increases with wider packet stripes as exhibited by the clustering about the line $y = x$. In Table II we summarize the statistics of the absolute error, i.e., the modulus of the difference between the inferred and actual link loss rates. This is just under 1% in the worst case, i.e., for the 2 packet stripe, and 0.63% in the best case, i.e., the 4 packet stripe. Thus, by exploiting the coalescence property, we have achieved a 40% reduction in absolute error, by simply increasing the the stripe length from two to four.

## V. Conclusions and Further Work

In this paper we have proposed a method of using end-to-end unicast probing to infer the loss characteristics of the network interior. The method relies on using collections of unicast probes, called stripes, dispatched back-to-back to different destinations, in order to mimic the effect of a notional multicast packet that followed the same path. We infer internal loss rates by applying an estimator developed for multicast inference to the unicast receiver traces. This estimator is unbiased when the transmissions of a stripe's probes on a given link are perfectly correlated. Imperfect correlations lead to bias, but we prove that this can be compensated for by using wider stripes, provided that the stripe transmissions obey a certain correlation property that we call coalescence. This is the property that successful transmission of a given packet in the stripe becomes more likely when more other packets from the stripe have been successfully transmitted.

Our network experiments show that for end-to-end transmission, correlations within stripes are very high, even perfect in some cases. Moreover, the coalescence property was found to hold in virtually all cases examined. Together these properties lead us to expect that inference from striped unicast probes will be effective in estimating link loss rates.

Our next step in network experimentation is to directly assess the method by performing corroborative measurements in the network interior. This entails taking measurements on paths over which probe traffic flows; then comparing actual loss rates with inferred loss rates on internal paths.

Currently, such corroboration is available to us only in simulation experiments. The `ns` simulations showed good agreement between inferred and actual loss rates; the typical error in these experiments was about 1% for the 2-packet stripe, falling to 0.63% when the stripe width was increased to 4.

Our next step in simulation will be to investigate the magnitude of these effects for systems with larger buffers and more diverse background traffic, which are more representative of actual networks.

In this paper we have concentrated on estimation of link probabilities for the first packet of a stripe. However, due to heterogeneity of loss along the stripe, such estimates may not be representative of typical packets, e.g., packets contained within a burst. Clearly, the present method could be extended, through use of other stripe templates, to estimate link probabilities for packet in positions other than the first. In the future we hope to increase the accuracy of inference by tuning the stripe properties to the burst structure observed in background traffic.

Finally, we remark that a number of other multicast-based estimators–namely those for delay distributions [15], for delay variances [7], and logical multicast topology [3]–have the potential to be adapted in the same manner as was done for loss estimators in this paper. We feel that our promising results on unicast-based loss estimation warrant extending the estimator to these other settings.

### *Acknowledgement*

We thank Ramon Caceres for his help with `ns`. Many thanks to Andrew Adams, his NIMI colleagues Matt Mathis and Jamshid Mahdavi, and the many NIMI volunteers who host NIMI measurement servers, for facilitating our Internet measurements.

## References

[1] A. Adams, T. Bu, R. Caceres, N.G. Duffield, T. Friedman, J. Horowitz, F. Lo Presti, S.B. Moon, V. Paxson, D. Towsley, The Use of End-to-End Multicast Measurements for Characterizing Internal Network Behavior, IEEE Communications Magazine., May 2000.

[2] R. Caceres, N.G. Duffield, J.Horowitz, D. Towsley, "Multicast-Based Inference of Network Internal Loss Characteristics", IEEE Trans. on Information Theory, 45: 2462–2480, 1999.

[3] R. Caceres, N.G. Duffield, J.Horowitz F. Lo Presti, D. Towsley, "Statistical Inference of Multicast Network Topology", in Proc. IEEE Conf. on Decision and Control, Phoenix, AZ, Dec. 1999.

[4] R. Carter, M. Crovella, 'Measuring bottleneck link-speed in packet-switched networks,' *Performance Evaluation*, 27&28, 1996.

[5] M. Coates, R. Nowak. "Network loss inference using unicast end-to-end measurement, to appear, *Proc. ITC Conf. IP Traffic, Modeling and Management*, Sept. 2000.

[6] A.B. Downey. "Using pathchar to estimate Internet link characteristics," *Proc. SIGCOMM'99* Sept. 1999.

[7] N.G. Duffield and F. Lo Presti, "Multicast Inference of Packet Delay Variance at Interior Network Links", in Proc. IEEE Infocom 2000, Tel Aviv, March 2000.

[8] S. Floyd. "Simulator tests." July 1995; revised May 1997. See http://www.aciri.org/floyd/papers/simtests.ps.Z

[9] S. Floyd and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance," IEEE/ACM Transactions on Networking, 1(4), August 1993.

[10] V. Jacobson, "Congestion Avoidance and Control," *Proc. SIGCOMM '88*, pp. 314-329, Aug. 1988.

[11] V. Jacobson, Pathchar - A Tool to Infer Characteristics of Internet paths. For more information see ftp://ftp.ee.lbl.gov/pathchar

[12] S. Keshav. " A control-theoretic approach to flow control," *Proc. SIGCOMM'91*, 3–15, Sept. 1991.

[13] K. Lai, M. Baker,"Measuring link bandwidths using a deterministic model of packet delay," *Proc. SIGCOMM 2000*, to appear.

[14] B.N. Levine, S. Paul, J.J. Garcia-Luna-Aceves, "Organizing multicast receivers deterministically according to packet-loss correlation", Preprint, University of California, Santa Cruz.

[15] F. Lo Presti, N.G. Duffield, J.Horowitz and D. Towsley, "Multicast-Based Inference of Network-Internal Delay Distributions", submitted for publication, September 1999.

[16] `mtrace` – Print multicast path from a source to a receiver. See ftp://ftp.parc.xerox.com/pub/net-research/ipmulti

[17] `ns` – Network Simulator. See http://www-mash.cs.berkeley.edu/ns/ns.html

[18] V. Paxson. "End-to-End Internet Packet Dynamics," *Proc. ACM SIGCOMM '97,* September 1997.

[19] V. Paxson, J. Mahdavi, A. Adams, M. Mathis, "An Architecture for Large-Scale Internet Measurement," IEEE Communications Magazine, Vol. 36, No. 8, pp. 48–54, August 1998.