# Network Loss Tomography Using Striped Unicast Probes

Nick Duffield, *Fellow, IEEE,* Francesco Lo Presti, Vern Paxson, Don Towsley, *Fellow, IEEE,*

***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 a multicast probe followed the same path 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. Using an M/M/1/K model for a link, we theoretically confirm this benefit for stripes. We also 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.**

***Index Terms*— End-to-end Measurement, Network Tomography, Packet Loss Rates, Estimation, Correlation**

## I. Introduction

### A. Motivation

As the Internet grows in size and diversity, its internal performance is becoming more difficult to measure. Any one organization has administrative access to only a fraction of the network's internal nodes, whereas commercial factors often prevent organizations from sharing internal performance data. *Multicast Inference of Network Characteristics* (MINC), avoids these problems by exploiting the inherent correlation in performance observed by multicast receivers to infer links level loss and delay statistics. These measurements require no participation by internal nodes besides forwarding packets.

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 reaches the receivers, it must have reached the node. By conditioning on arrival at a descendent, we can determine the probability of successful transmission to and beyond the given node. Inference algorithms are given in [3] for loss, [23] for delay distributions, [11] for delay variances, and [10] for inferring the logical multicast tree topology itself.

However, these methods suffer 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 may differ significantly from that observed by unicast packets, while unicast traffic constitutes the bulk of Internet traffic. This motivates inference methods for network internal performance that are based on end-to-end unicast measurements. A challenge here is that unicast does not exhibit the packet-level correlation at different receivers that is possessed by multicast.

### B. Contribution

In this paper we adapt the multicast inference techniques proposed in [3] 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 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. We summarize how multicast inference works; see [3] for further details.

Consider the tree shown in Figure 1. Suppose multicast probes are dispatched from a source at node $0$ towards the leaves at nodes $\mathsf{l}$ and $\mathsf{r}$. The problem is to estimate the link transmission probabilities $\alpha_c$, $\alpha_\mathsf{l}$ and $\alpha_\mathsf{r}$ from end-to-end measurements. The expected proportion of probes transmitted successfully to node $\mathsf{l}$ is $A_\mathsf{l} = \alpha_\mathsf{c}\alpha_\mathsf{l}$, and to node $\mathsf{r}$ is $A_\mathsf{r} = \alpha_\mathsf{c}\alpha_\mathsf{r}$. These relations would also hold for unicast packets sent independently to the leaves. But with multicast, the expected proportion of probes that reach both nodes is $A' = \alpha_\mathsf{c}\alpha_\mathsf{l}\alpha_\mathsf{r}$. The foregoing three relations can be used to recover the link transmission probabilities $\alpha_\mathsf{c}, \alpha_\mathsf{l}$ and $\alpha_\mathsf{r}$ in terms of the end-to-end transmission probabilities $A_\mathsf{l}$, $A_\mathsf{r}$ and $A'$:

$$\alpha_\mathsf{c} = A_\mathsf{l}A_\mathsf{r}/A', \quad \alpha_\mathsf{l} = A'/A_\mathsf{r}, \quad \alpha_\mathsf{r} = A'/A_\mathsf{l} \qquad (1)$$

Substituting the corresponding *measured* end-to-end transmission probabilities $\widehat{A}_\mathsf{l}, \widehat{A}_\mathsf{r}$, and $\widehat{A}'$ into these relations yields estimates of the link probabilities. These estimates can be shown to be maximum likelihood estimators; moreover, the whole approach generalizes to trees of arbitrary topology.

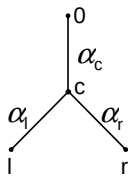


Fig. 1. Two-leaf tree, with marginal link transmission probabilities

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

The central idea in this paper 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 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 experience very similar behavior when they traverse common portions of their paths. 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. The packet loss and delays on a link would be perfectly correlated within a stripe; the composite probe would have the same statistical properties as an imaginary multicast packet that followed the same source tree. In this case the methods of [3], [11], [23] could be applied immediately to infer the link loss and delay statistics of the logical source tree.

However, correlations within stripes may be less than perfect in practice. 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) [14] is another mechanism by which packet loss may become decorrelated. However, this mechanism is seldom used in the Internet. 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) developing a theory with which to develop measurement procedures to reduce the impact of imperfect correlations;
(ii) finding the impact of these procedures on the accuracy of inference methods that assume perfect correlations;
(iii) determining the magnitude of imperfect correlations through experiments on real networks;
(iv) verifying the accuracy of the approach in simulations.

We extend the packet loss model of [3] by incorporating an additional parameter for each link that describes the correlation of loss between different packets of the same stripe. 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. Calculations show that the error in using the loss estimator from [3] is small provided that the conditional probability of the *loss* of one packet in the stripe given the *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 for single paths through measurement.

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

The rest of the paper is as follows. Section II formulates the stripe method for binary then general trees. We specify a family of striping methods. We state the assumption on packet correlations within stripes, construct a hierarchy amongst the various striping methods based on the correction made to the bias caused by imperfect correlations. Section III shows that the correlation assumption is satisfied when the common path is modeled by an M/M/1/K queue. We also present some numerical calculations that show the benefits for estimator accuracy of conditioning in certain types of longer stripes.

We use two experimental approaches to evaluate the proposed method. Section IV uses end-to-end measurement on the National Internet Measurement Infrastructure (NIMI) [27] 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. We estimated the likely accuracy of stripe-based inference in the actual network.

We support this work in Section V using network level simulation with `ns` [25]. By instrumenting the simulation we can trace the behavior of packets in the network interior. This allows us to study the correlation properties of packets within stripes on links (as opposed to paths), and to compare the inferred link loss rates with actual link loss rates. For the most accurate choice of striping method the typical absolute error in loss rate inference is below 1%. We conclude in Section VI. Proofs of the Theorems are in Section VII.

### C. Related Work

Several tools and methodologies can 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 [18], [20] 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. The idea was embodied in a number of tools; see [5], [21], [26]. Although these methods focus on bandwidth estimation, they are based on the same idea, namely to send packet pairs (or stripes) so as to introduce correlation in a controlled manner. These methodologies have also been applied successfully to estimate cross traffic characteristics; see [17], [29]. Packet pair probing has been applied to the related problem is to determine whether two flows share a common congestion bottleneck; see [16], [30]. This problem is substantially different than that of estimating link loss *rates*.

[6] uses 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 transmission probabilities of all links in the tree. Conditional transmission probabilities

are introduced as unknown nuisance variables. The inference of the link transmission probabilities is formulated as a maximum likelihood estimation problem which is solved using the Expectation Maximization (EM) algorithm.

Our approach differs from [6]: we consider a more general form of striping which yields significantly higher correlation. Thus we are able to continue to rely on the maximum likelihood estimates derived for the multicast case for which we have closed form expressions which are readily computable without need for iteration. In network simulations, our methods suffered an absolute error of about 1% in estimating link loss rates in a 7 receiver tree, down to about 0.3% in larger trees, errors being smaller for longer stripes. Results with the EM-based approach appear to be comparable, with worst case absolute error of about 2% in estimating loss; see [31].

Potentially all other multicast based inference techniques, the estimators for delay distributions [23], for delay variances [11], and logical multicast topology [10] can be adapted to unicast measurements. Similar developments have been performed for the EM-based approaches; see [7].

Finally, `pathchar` [8], [19] 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 fixed 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 $j$ and then $k$ without encountering another element of $V$ in between. We will sometimes refer to link $(j,k) \in L$ simply as link $k$. For $k \neq 0$, $f(k)$ denotes the parent of $k$. 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 $r_1, \ldots, r_{d_0}$. For each $k \in V$ we will denote $D(k) \subset D_0 = \{1, \ldots, d_0\}$ the set of packets that transit across link $k$ en route to their destinations. We will refer to $D(k)$ as the substripe to node $k$. We describe the progress of the stripe in $\mathcal{T}$ by the variables $X_k(d)$, $d \in D(k)$, 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.

It is useful to have a notation describing composite events at sets of receivers. For $D \subset D_0$ define the binary variable

$$Z_D = \prod_{d \in D} X_{r_d}(d). \tag{2}$$

$Z_D = 1$ if all packets in $D$ reach their destinations, and $0$ otherwise. It is 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 $D \subset D(k)$, $k \in V$, let $\alpha_k(D)$ denote the probability that all packets in $D$ are transmitted across link $k$ 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 $d \in D(k)$. We will refer to the various $\alpha_k$ as the link $k$ transmission probabilities and to the complement $1 - \alpha_k$ as the link $k$ loss probabilities. For disjoint subsets $D, D' \subset D(k)$, let $\beta_k(D|D')$ denote the conditional probability that packets in $D$ are successfully transmitted across link $k$ to node $k$, given that those in $D'$ are also successfully transmitted across link $k$, all packets having reached the parent node $f(k)$. This is expressed in terms of the probabilities $\alpha_k$ as

$$\beta_k(D|D') = \alpha_k(D \cup D')/\alpha_k(D'). \tag{3}$$

The $\beta_k$ are the link $k$ conditional transmission probabilities.

With perfect correlations the $\beta_k$ would be 1. The multicast loss model of [3] is statistically equivalent to the special case $\beta_k(D|D') = 1$ and hence $\alpha_k(d)$ all equal some $\alpha_k$.

For a given link and substripe width, the structure of the probabilities $\alpha, \beta$ should depend on the times between successive packets. If the packets are widely separated, 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 case the marginal transmission probabilities will depend on the position of a packet within a stripe, particularly when the stripe width is not negligible compared with buffer sizes. For each link $k$, we focus on estimating the transmission probability $\alpha_k(1)$ of the first probe in the substripe traversing $k$. 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 [3] under imperfect correlations. We start with the two-leaf tree shown in Figure 1, having leaf nodes $\mathsf{l}$ and $\mathsf{r}$ with common parent $\mathsf{c}$ whose own parent is the root $0$. Consider the binary stripe $\langle \mathsf{l}, \mathsf{r} \rangle$. The link probabilities are related to the probabilities of leaf events as follows:

$$\frac{\mathsf{E}Z_1\mathsf{E}Z_2}{\mathsf{E}Z_{12}} = \frac{\alpha_{\mathsf{c}}}{\beta_{\mathsf{c}}(1|2)},\ \frac{\mathsf{E}Z_{12}}{\mathsf{E}Z_2} = \alpha_{\mathsf{l}}\beta_{\mathsf{c}}(1|2),\ \frac{\mathsf{E}Z_{12}}{\mathsf{E}Z_1} = \alpha_{\mathsf{r}}\beta_{\mathsf{c}}(2|1),$$

where $Z_D$ is as defined in (2) and $\alpha_k$, $k \in \{\mathsf{c}, \mathsf{l}, \mathsf{r}\}$, is the transmission probability of the first probe in the substripe traversing link $k$, *i.e.*, $\alpha_{\mathsf{c}} = \alpha_{\mathsf{c}}(1)$, $\alpha_{\mathsf{l}} = \alpha_{\mathsf{l}}(1)$ and $\alpha_{\mathsf{r}} = \alpha_{\mathsf{r}}(2)$. Expressions (4) are obtained by expanding $\mathsf{E}Z_D$, e.g., $\mathsf{E}Z_{12} =$

$\alpha_{\mathsf{c}}(12)\alpha_{\mathsf{l}}(1)\alpha_{\mathsf{r}}(2) = \alpha_{\mathsf{c}}(2)\beta_{\mathsf{c}}(1|2)\alpha_{\mathsf{l}}(1)\alpha_{\mathsf{r}}(2)$, with similar expressions for $\mathsf{E}Z_1$ and $\mathsf{E}Z_2$. With perfect correlations, $\beta_{\mathsf{c}} = 1$, the $\alpha$ may be recovered directly from the leaf probabilities.

These expressions are used to estimate the $\alpha$ from the leaf events $Z^{(i)}$ associated with multiple identical stripes $i = 1, 2, \ldots n$. We replace each expectation in (4) by the corresponding empirical mean, defined here in general:

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

Taking $\beta_{\mathsf{c}} = 1$ then yields the estimates

$$\widehat{\alpha}_{\mathsf{c}} = \widetilde{Z}_1\widetilde{Z}_2/\widetilde{Z}_{12},\ \widehat{\alpha}_{\mathsf{l}} = \widetilde{Z}_{12}/\widetilde{Z}_2,\ \widehat{\alpha}_{\mathsf{r}} = \widetilde{Z}_{12}/\widetilde{Z}_1. \tag{5}$$

This is the estimator from [3] applied to the two-leaf tree.

With imperfect correlations, $\beta_{\mathsf{c}}$ cannot be recovered independently from the leaf expectations. The model is not identifiable; this was observed in [6]. Since $\beta_{\mathsf{c}} \leq 1$, estimation via (5) is biased, overestimating $\alpha_{\mathsf{c}}$ and underestimating $\alpha_{\mathsf{l}}, \alpha_{\mathsf{r}}$.

*1) The complementary stripe* $\langle \mathsf{r}, \mathsf{l} \rangle$*:* We now consider the complementary binary stripe $\langle \mathsf{r}, \mathsf{l} \rangle$ obtained by exchanging the order of packet destinations. The link probabilities are now related to the probabilities of leaf events as follows:

$$\frac{\mathsf{E}Z_1\mathsf{E}Z_2}{\mathsf{E}Z_{12}} = \frac{\alpha_{\mathsf{c}}}{\beta_{\mathsf{c}}(1|2)},\ \frac{\mathsf{E}Z_{12}}{\mathsf{E}Z_2} = \alpha_{\mathsf{l}}\beta_{\mathsf{c}}(2|1),\ \frac{\mathsf{E}Z_{12}}{\mathsf{E}Z_1} = \alpha_{\mathsf{r}}\beta_{\mathsf{c}}(1|2).$$

To estimate $\alpha$ from measurements, we take the expectations in (6). With $\beta_{\mathsf{c}} = 1$, we obtain again (5). Note that, however, despite having identical expressions, the estimators are different. The conditional probabilities in the second and third expressions in (4) and (6) are exchanged as a consequence of the the destinations of the two probes being inverted. With perfect correlation, either stripe yields unbiased estimates. With imperfect correlation, both estimates are biased, the bias being asymmetrical for $\widehat{\alpha}_{\mathsf{l}}$ and $\widehat{\alpha}_{\mathsf{r}}$ depending on the relative values of $\beta_{\mathsf{c}}(1|2)$ and $\beta_{\mathsf{c}}(2|1)$. Queueing analysis (Section III) and network experiments (Section IV) suggest that $\beta_{\mathsf{c}}(1|2)/\beta_{\mathsf{c}}(2|1) \geq 1$ with the ratio increasing as the marginal loss probability does. Therefore we expect $\widehat{\alpha}_{\mathsf{r}}$ to have a larger bias than $\widehat{\alpha}_{\mathsf{l}}$ (and $\widehat{\alpha}_{\mathsf{c}}$) for the stripe $\langle \mathsf{l}, \mathsf{r} \rangle$ and the reverse for the stripe $\langle \mathsf{r}, \mathsf{l} \rangle$. To avoid this inherent asymmetry in estimator accuracy, we use both types of stripes and retain the estimates from each set of results. Their bias will depend only on the conditional probability closer to 1, *i.e.*, $\beta_{\mathsf{c}}(1|2)$.

### C. Enhancing Stripe Correlations

Uncertainty in $\beta(1|2)$ undermines confidence in using (5) directly. We propose a modified stripe for which the effective value of $\beta$ is closer to 1. For the stripe $\langle \mathsf{l}, \mathsf{r} \rangle$ with perfect correlations, $\mathsf{E}Z_{12}/\mathsf{E}Z_2$ (the conditional probability for the first packet of the stripe to reach $\mathsf{l}$ given that its second packet reaches $\mathsf{c}$) is actually equal to the probability of transmission of a packet along the link $(\mathsf{c}, \mathsf{l})$, conditional upon reaching $\mathsf{c}$. This is because packet 2 must have been present at $\mathsf{c}$ if present at $\mathsf{r}$. With imperfect correlations, packet 1 may not have been also present at $\mathsf{c}$, leading to underestimation of $\alpha_{\mathsf{l}}$. Our remedy is to use longer stripes, conditioning on an event at $\mathsf{r}$ which makes it more likely that packet 1 was present at $\mathsf{c}$.

The simplest example is the **three-packet stripe** $\langle \mathsf{l}, \mathsf{r}, \mathsf{r} \rangle$. Provided that transmission of packets within the stripe is strongly correlated (as specified in Definition 1 below) it should be more likely that packet 1 reaches $\mathsf{c}$, upon reception of packets 2 and 3 at receiver $\mathsf{r}$, rather with than reception of packet 2 alone. Conditioning on reception of packets 2 and 3, the analogs of the first and second relations in (4) are

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

The parameters $\alpha_{\mathsf{c}}$ and $\alpha_{\mathsf{l}}$ are estimated by $\widetilde{Z}_1\widetilde{Z}_{23}/\widetilde{Z}_{123}$ $\widetilde{Z}_{123}/\widetilde{Z}_{23}$ respectively; $\alpha_{\mathsf{r}}$ can be estimated similarly using the complementary stripe $\langle \mathsf{r}, \mathsf{l}, \mathsf{l} \rangle$. Comparing with (5), these estimates introduce less bias than those from two-packet stripes provided that $\beta_{\mathsf{c}}(1|2,3) > \beta_{\mathsf{c}}(1|2)$. This is the case provided transmissions 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 states that a set of packets is more likely to be transmitted on a link after other packets from the stripe have been transmitted over that link. Coalescence real network traffic is investigated in Section IV. Conditioning with more packets, the effect is to decrease the estimate of $\alpha_{\mathsf{c}}$ and to increase the estimate of $\alpha_{\mathsf{l}}$ or $\alpha_{\mathsf{r}}$. Thus, we can counteract the bias in the two-leaf stripe (see (4)) with wider stripes.

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

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

(8) says that extending the stripe reduces the estimate of the transmission rate $\alpha_{\mathsf{c}}$ and so counteracts the bias due to $\beta_{\mathsf{c}} < 1$.

Example: the 4-packet stripe: Theorem 1 suggests that we can further reduce bias by lengthening the stripe length. Consider, for instance, the stripe $\langle \mathsf{l}, \mathsf{r}, \mathsf{r}, \mathsf{r} \rangle$ and compare its estimation properties with those of its substripes $\langle \mathsf{l}, \mathsf{r}, \mathsf{r} \rangle$ and $\langle \mathsf{l}, \mathsf{r} \rangle$. By Theorem 1 we have the following ordering:

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

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. 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 [3] for nodes with arbitrary numbers of offspring in order to estimate the link probabilities. A potential problem with is that the statistical properties of stripes may not reflect those of general traffic if their width is not negligible compared with

buffer sizes. Instead, here we focus on combining inferences from fixed-width stripe measurements on embedded subtrees.

Consider an arbitrary tree with leaf set $R$. For each node $k \in V \setminus R$ 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 $k$ and $M(k)$ a subset of $Q(k)$ such that $(i,j) \in M(k)$ iff $(j,i) \in M(k)$. For each $(i,j) \in M(k)$, consider the embedded two-leaf binary tree spanned by the nodes $0, k, i, j$. 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 $(\mathsf{r}_{i_1}, \mathsf{r}_{i_2})$ in $M(k)$ we form a stripe that sends packets in positions in $D_1$ to $\mathsf{r}_{i_1}$ and packets in positions in $D_2$ to $\mathsf{r}_{i_2}$. More formally, this is the stripe $S(\mathsf{r}_{i_1}, \mathsf{r}_{i_2}) = \langle r_1, \ldots, r_{d_0} \rangle$ where $r_d = \mathsf{r}_{i_\ell}$ when $d \in D_\ell$.

The relation between the leaf and 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 $A_k(D) = \prod_{j \succeq k} \alpha_j(D)$ and $B_k(D_1|D_2) = \prod_{j \succeq k} \beta_j(D_1|D_2)$. Henceforth, we omit the dependence on $D$, $D_1$, and $D_2$ when the context is clear. Here, we consider the same type of stripes described in Section II: for the ordered pair $(\mathsf{r}_{i_1}, \mathsf{r}_{i_2})$, we assume $D_1 = \{1\}$ and $D_2 = \{2, \ldots, d_0\}$, *i.e.*, the stripe $S(\mathsf{r}_{i_1}, \mathsf{r}_{i_2}) = \langle \mathsf{r}_{i_1}, \mathsf{r}_{i_2}, \ldots, \mathsf{r}_{i_2} \rangle$. The pair $(\mathsf{r}_{i_2}, \mathsf{r}_{i_1})$ corresponds to the complementary stripe $S(\mathsf{r}_{i_2}, \mathsf{r}_{i_1}) = \langle \mathsf{r}_{i_2}, \mathsf{r}_{i_1}, \ldots, \mathsf{r}_{i_1} \rangle$ sent down the same subtree. For each non-leaf and non-root node $k$ and each pair $(i,j) \in M(k)$, the measurements with $n$ stripes of type $S(i,j)$ give rise to an estimate of $A_k = \prod_{j \succeq k} \alpha_j$

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

In this paper we use arithmetic mean of estimates

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

For each leaf node $k$, 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 the quotients

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

### *E. Asymptotic behavior of loss estimates*

*Theorem 2:* For $k \neq 0$, $\sqrt{n} \cdot (\widehat{\alpha}_k - \alpha_k - m_k)$ converges, as $n \to \infty$, to a mean zero Gaussian r.v. of variance $\sigma_k^2$, where

$$m_k = \begin{cases} \frac{\alpha_k}{\beta_k}(1 - \beta_k) & k \notin R \\ \alpha_k(B_{f(k)} - 1) & k \in R. \end{cases} \tag{14}$$

Theorem 2 shows that with imperfect correlation, $\widehat{\alpha}_k$, $k \neq 0$, computed via (13), are biased. We define the estimator bias as $b_k := |\mathsf{E}[\widehat{\alpha}_k - \alpha_k]|$, $k \neq 0$. For large $n$ we can use the approximation $b_k \approx |m_k|$. From (14), the estimator bias depends on the position of the link in the tree. The bias for a leaf link depends on the conditional transmission probabilities along the entire end-to-end path from the source. Since these conditional probabilities typically decrease with path length, the bias should grow with the size of the tree. The estimator bias of a non-leaf link, instead, only depends on the transmission probabilities of that link, not the tree size.

The analysis of the asymptotic variance $\sigma_k^2$ can be performed along the same lines used for that for multicast inference [3]. Here we will focus for simplicity on the regime in which all loss rates $\overline{\alpha}_k = 1 - \alpha_k$ are close to zero. In this regime it is not difficult to show that

$$\sigma_k^2 = \frac{s(k)}{\#M(k)} + \frac{s(f(k))}{\#M(f(k))} + O(\|\overline{\alpha}\|^2) \tag{15}$$

where $\|\overline{\alpha}\| = \max_{k \in V} \overline{\alpha}_k$ and $s(k) = \sum_{j \succeq k} \overline{\alpha}_k$ is the loss rate along the path from 0 to node $k$ (it is easy to verify that in this regime $A(k) = 1 - s(k) + O(\|\overline{\alpha}\|^2)$). To leading order, $\sigma_k^2$ is proportional to the loss rate from the source to node $k$, and inversely proportional to the number of subtrees used in estimating $\widehat{A}_k$ and $\widehat{A}_{f(k)}$. Thus the estimator variance depends on the topology and size of the tree and grows with the distance from the probe source. This differs from the analogous result for multicast inference (see [3]), where to leading order the variance is independent of topology.

### *F. Measurement Approaches*

Inference for general logical tree works by combining estimates from measurements on embedded 2 receiver subtrees. In the *exhaustive striping strategy* measurements are taken across all binary subtrees, i.e., by taking $M(k) = Q(k)$. In the *minimal striping strategy* measurements are limited to a single subtree passing through each node $k$, taking, e.g., $M(k) = \{(i,j),(j,i)\}$ for some receivers $i, j$ depending on $k$. (It can be shown that measurement must be made on at least one such subtree per node in order to estimate all the link probabilities; see [9]). The minimal strategy has several advantages. First, it scales better: if we fix the number of stripes sent down each subtree, it requires a total number of stripes which grows linearly with the number of nodes while a complete set of measurements requires a number of probes proportional to the square of the number of nodes. Second, it provides estimates with lower variance. To see this, we compare the asymptotic estimator variance $\sigma_k^2$ for a fixed total number of stripes $m$. Assume a binary topology of depth $d$; each type stripe is then transmitted $\#R(\#R-1)m/2$ times in the complete case and $\#R\,m/2$ in the single subtree case, with $\#R = 2^{d-1}$. Then in the asymptotic regime of small loss

$$\frac{\mathsf{Var}[\widehat{\alpha}_k]_{single}}{\mathsf{Var}[\widehat{\alpha}_k]_{complete}} = \frac{1}{2^{d(k)-1}} \frac{s(k) + s(f(k))}{2s(k) + s(f(k))} + O(\|\overline{\alpha}\|^2) < 1$$

for $k \neq 0$, where $d(k)$ denotes the depth of node $k$ and $\mathsf{Var}[\widehat{\alpha}_k]_{complete}$ and $\mathsf{Var}[\widehat{\alpha}_k]_{single}$ denote the variances in the two cases. Thus, for each $k$, the single subtree approach always yields a smaller variance with the ratio decreasing exponentially with the depth of $k$ in the tree. So reduction of measurement subtrees is more than compensated by the larger number of probes sent down each subtree.

### G. Sampling and Statistical Issues

We now make two further observations of the statistical implications of using the stripe approach. First, network characteristics may not be uniform across a stripe e.g., if stripe width is comparable in size to that of a buffer. The 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.

Second, there is a phenomenon during TCP slow start, in which every other or every third packet being lost due to specific buffer-filling patterns; see Figure 2 of [13]. These may impart particular loss patterns on the elements of a stripe.

## III. Coalescence in the M/M/1/K Queue

In this section we analyze coalescence in the context of an M/M/1/K queue. While our proof does not extend to general queuing systems, the analysis of this simple case provides useful insights on probe transmission characteristics.

We model a network node as an M/M/1/K queue with a Drop Tail discard policy. The queue is offered background traffic according to a Poisson process with rate $\lambda_b$. The queue also receives a stream of (non Poisson) probe traffic comprising packets stripes. We assume the interarrival times between stripes are i.i.d. with mean $1/\lambda_s$, $\lambda_s \ll \lambda_b$; each stripe comprises $d_0$ probes with exponential interarrival time of mean $1/\lambda_p$. Last, we assume all packet service times are i.i.d. exponential random variables with mean $1/\mu$.

To study the probe transmission probabilities we analyze the transient behavior of the queue in the interval of time from the arrival of the first probe of a stripe until the arrival of the last. the queue is offered aggregate Poisson traffic with rate $\lambda_a = \lambda_p + \lambda_b$. The number of packets found in the queue by successive arrivals (either background or probe traffic) is then a Markov chain with one step transition probability matrix

$$P_a = \begin{pmatrix} 1-a_0 & a_0 & 0 & 0 & \cdots & 0 \\ 1-a_0-a_1 & a_1 & a_0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & 0 \\ 1-\sum_{j=0}^{K} a_j & a_K & a_{K-1} & a_{K-2} & \cdots & a_0 \end{pmatrix}$$

where $a_j$ is the probability that $j$ packets left the queue between two consecutive arrivals. For exponentially distributed interarrival and service times one sees the number of departure between arrivals is geometrically distributed with success probability $\frac{\lambda_a}{\lambda_a+\mu}$; thus, $a_j = \frac{\lambda_a \mu^j}{(\lambda_a+\mu)^{j+1}}$, $j = 0, \ldots, K$.

Consider now the number of packets $N_d$, $d = 1, \ldots, d_0$, found in the queue by the $d$-th probe in a stripe. It is easy to see that $N_d$ is also a Markov chain. Its transition probability matrix $P_p$ can be computed from $P_a$ by conditioning on the number of background arrivals between two consecutive probe arrivals. By observing that the number of background arrivals between two consecutive probe arrivals is geometrically distributed with parameter $\frac{\lambda_b}{\lambda_b+\lambda_p}$, it immediately follows that

$$\begin{aligned} P_p &= \sum_{j=0}^{\infty} P_a^{j+1} \frac{\lambda_p \lambda_b^j}{(\lambda_p+\lambda_b)^{j+1}} \\ &= \frac{\lambda_p}{\lambda_b}\left[\left(I - \frac{\lambda_b}{\lambda_b+\lambda_p} P_a\right)^{-1} - I\right] \end{aligned} \tag{16}$$

where $I$ denotes the $(K+2)\times(K+2)$ identity matrix.

Let $\pi_j = (\pi_j(0), \ldots, \pi_j(K+1))$, $\pi_j(i) = \mathsf{P}[N_j = i]$, $j = 1, \ldots, d_0$, denote the vector of state probabilities seen by the $j$-th probe of a stripe. For $j = 2, \ldots, d_0$, $\pi_j = \pi_1 \cdot P_p^j$. Since $\lambda_s \ll \lambda_b$, we can assume that the queue reaches its steady state between two consecutive stripes so that $\pi(1)$ and, hence, $\pi_1$ is given by the steady state distribution of the queue fed by the background traffic only. Then, standard results for the M/M/1/K queue yield $\pi_1(i) = \rho^i \frac{1-\rho}{1-\rho^{K+2}}$, where $\rho = \lambda_b/\mu$ is the offered background traffic.

Loss occurs when a probe packet finds the system full. The marginal probability of successful transmission for the $j$-th probe within a stripe is then $\alpha(j) = \sum_{i=0}^{K} \pi_j(i)$. For $D = \{d_1, \ldots, d_m\} \subseteq D_0$, the probability of joint successful transmission of all probes in $D$, $\alpha(D)$, is $\alpha(D) = \mathsf{P}[N_{d_l} \leq K, l = 1, \ldots, m]$. To compute $\alpha(D)$, we write $\alpha(D) = \sum_{i=0}^{K+1} \phi(i)$, where we denote $\phi(i) = \mathsf{P}[N_{d_l} \leq K, l = 1, \ldots, m, N_1 = i]$. It is easy to verify that $\phi = (\phi(0), \ldots, \phi(K+1))$ obeys

$$\phi = \widetilde{\pi}_1 \cdot \widetilde{P}_{p,2} \cdots \widetilde{P}_{p,d_m}, \tag{17}$$

where

$$\widetilde{\pi}_1 = \begin{cases} \pi_1 & \text{if } 1 \in D \\ [\pi_1]_{K+1} & \text{if } 1 \notin D \end{cases} \tag{18}$$

$$\widetilde{P}_{p,j} = \begin{cases} P_p & \text{if } j \in D \\ [P_p]_{K+1} & \text{if } j \notin D \end{cases} \tag{19}$$

and $[.]_{K+1}$ denotes the same vector (matrix) with the $K+1$-th element (row and column) replaced with a zero. The conditional probabilities are obtained by the appropriate quotients.

### A. Structure of $\alpha$ and $\beta$ and Coalescence Property

For this model, we can establish the structure of the probabilities $\alpha$ and $\beta$ by studying the stochastic order relations among the $N_d$, $d \in D_0$, as detailed in Section VII.

*Theorem 3:* In a M/M/1/K queue the stripe transmission is coalescent. Moreover,

(i) $\alpha(1) \geq \alpha(2) \geq \cdots \geq \alpha(d_0)$;
(ii) $\beta(1|2) \geq \beta(2|1)$.

Theorem 3 establishes the coalescence property for the M/M/1/K queue. In addition, (i) states that the marginal transmission probability decreases with the packet position in the stripe. Intuitively, if probes arrive in rapid succession to a finite buffer so that there is no arrival or service completion in the interval of time between the first and last probe, probe loss occurs when there is insufficient buffer space to accommodate the entire stripe; under Drop Tail discard, the last probes are those more likely to experience loss. Moreover, (ii) shows that, of the two conditional probabilities affecting the behavior of the two packet stripe, $\beta(1|2)$ is closer to one.

### B. Numerical Examples

We now evaluate the M/M/1/K model to illustrate the dependence of the transmission probabilities on system parameters, and the consequences for estimator accuracy.

*1) Structure of the Transmission Probabilities:* In Figure 2 we plot $\alpha$ and $\beta$ as function of the offered background traffic load $\rho = \lambda_b/\mu$, for different probe arrival rates when the buffer size is $K = 50$. We take $\mu = 1$, so that $\lambda_b = \rho$. The marginal transmission probabilities decrease with the packet position in the stripe, the difference becoming more marked as $\rho$ increases, i.e. when the link is congested.

The main observations concern the dependence on the probe arrival rate $\lambda_p$. For $\lambda_p = 1000$, probes arrive practically back to back since the probe arrival rate is three orders of magnitude higher than both background traffic arrival and service rates. The conditional probabilities show that loss of different probes in the stripe is highly correlated: $\beta(1|2)$, $\beta(1|23)$ and $\beta(1|234)$ are practically equal to one irrespective of the traffic load. The conditional probability $\beta(2|1)$ is also very close to $\alpha(1)$, although smaller than $\beta(1|2)$, especially as load increases.

As interarrival times increase, probe correlation decreases. When $\lambda_p = 0.1$, probe interarrival times are large and we expect the congestion events experienced by probes to be practically independent. All marginal and conditional probabilities are close to $\alpha(1)$ and lengthening the stripe leads to only a small increase in the conditional probabilities.

For intermediate probe arrival rates (not shown) conditional probabilities are reduced, with $\beta(1|2)$ decreasing fastest as the load $\rho$ increases. This indicates that we can counteract the decrease in probe correlation by using longer stripes.

In general, we expect $\beta(1|2)$ to be smaller than $\beta(2|1)$ because, while it is likely that the first probe of a stripe will find a non-full queue if succeeding probes do also, the reverse is not true. For example, when a probe occupies the last available position in the queue all the successive probes are lost at least until the first probe is processed. We also considered other buffer sizes and observed the relative behavior of $\alpha$ and $\beta$ to be insensitive to the buffer size: increasing (decreasing) $K$ only results in shifting the onset of congestion to a higher (lower) load.

*2) Estimator Bias:* We now illustrate the dependence of the inference method accuracy on the stripe structure. The setting is the two-leaf tree in Figure 1 with an M/M/1/K queue with buffer $K = 50$ at the common link; in this topology bias arises only through imperfect correlations at the common link. To quantify the accuracy of the estimates we compare the estimated loss probability of the common link $1 - \widehat{\alpha}_c$ with the actual loss probability $1 - \alpha_c$ by computing the estimator relative bias as the ratio $\frac{\mathsf{E}[\widehat{\alpha}_c - \alpha_c]}{1-\alpha_c} = \frac{\alpha_c}{\beta_c}\frac{1-\beta_c}{1-\alpha_c}$. In Table I we display the relative bias, expressed as a percentage, for different stripe widths $w$, probe interarrival rates and link loads, together with the conditional probability $\beta(1|2\ldots w)$.

The main observation is the dramatic decrease in the bias for longer stripes and higher probe interarrival rates. To ensure maximum correlation, probes should be transmitted back to back: in these examples, for $\lambda_p = 1000$, the bias is practically zero. In practice, this may not be sufficient since probes can be spaced apart, resulting in a smaller value of $\lambda_p$, as a result of traversing a bottleneck link. This could affect the accuracy of the estimates. Nevertheless, it is possible to counteract the correlation decrease by using longer stripes. For $\lambda_p = 10$, the increase of the stripe width from 2 to 4 reduces the bias, which would be otherwise as high as 10%, to below 1%. Bias is affected, but to a smaller extent, by the traffic load; because of the smaller value of $\beta$, bias increases with $\rho$, the difference being more significant for longer stripes.

| $\rho = 1$ | stripe width $w$ | | | | | |
|---|---|---|---|---|---|---|
| | 2 | | 3 | | 4 | |
| $\lambda_p = 1000$ | 0.10% | *1* | 0% | *1* | 0% | *1* |
| $\lambda_p = 100$ | 1.00% | *0.9998* | 0.02% | *1* | 0% | *1* |
| $\lambda_p = 10$ | 8.55% | *0.9983* | 1.35% | *0.9997* | 0.26% | *0.9999* |
| $\lambda_p = 1$ | 39.0% | *0.9924* | 22.0% | *0.9957* | 13.8% | *0.9973* |

TABLE I

ESTIMATOR BIAS AND CONDITIONAL PROBABILITIES. BIAS OF THE ESTIMATES OF THE LINK LOSS PERCENTAGE IN STRIPES OF WIDTH $w = 2$ TO 6 FOR DIFFERENT VALUES OF PROBE INTERARRIVAL RATE, TOGETHER WITH *conditional probabilities* $\beta(1|2\ldots w-1)$

## IV. NETWORK EXPERIMENTS

The techniques described in Section II rely on conditional probabilities of packet transmission within stripes being close to one, and the coalescence property in order to produce low bias estimators. 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, 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.

### A. Measurement Infrastructure and Datasets

We conducted the experiments using the National Internet Measurement Infrastructure (NIMI) [27]. 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. `zing` was extended to transmit unicast stripes to multiple destinations with minimal spacing between packets. This is done 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. The packet size was 60 bytes. A key point is that all packets in a stripe are sent to the same destination, with the goal being to assess the conditional loss probability and coalescence properties of paths.

A total of 83 successful measurements were made between 35 NIMI sites, each measurement being recorded at both sender and receiver. The measurement transmissions were of three types (i) 100,000 flights of stripes of 3 packets, with separations exponentially distributed with a mean of 100 msec; (ii) 20,000 flights of stripes of 3 packets, separated by a mean of 500 msec; (iii) 6000 flights of stripes of 10 packets separated by a mean of 100 msec; (iv) 10,000 flights of stripes


Fig. 2. TRANSMISSION PROBABILITIES: M/M/1/K QUEUE. Transmission probabilities $\alpha$ and $\beta$ as function of the offered background traffic load $\rho = \lambda_b/\mu$ ($\mu = 1$) for different values of $\lambda_p$. Buffer length is $K = 50$.

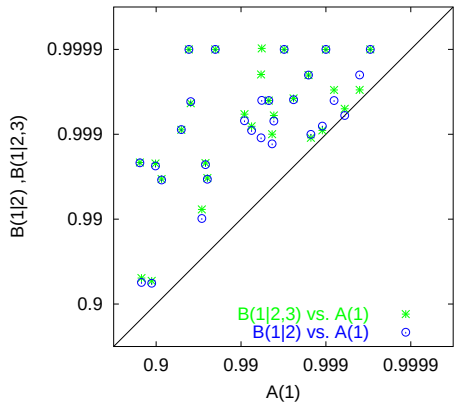


Fig. 3. SCATTER PLOT OF TRANSMISSION PROBABILITIES IN 38 NETWORK EXPERIMENTS. Conditional vs. marginal end-to-end transmission probabilities. Probabilities for 3-packet stripes mostly meet or exceed those for 2-packet stripes.

of 10 packets, separated by a mean of 300 msec. In the latter two cases, we also counted the first 3 packets in each stripe as another dataset of 3-packet stripes. All measurements were made at either 2PM EDT (a busy period) or 2AM EDT (a period of light load). The worst case average probe load was 100 packets per second. There was no noticeable change in measured transmission rates as we varied the inter-stripe spacing from 100 msec to 500 msec.

### B. Measured Transmission Probabilities

*a) Marginal Probabilities.:* Packet loss rate ranged between zero and $14\%$. Of the 83 traces, 13 exhibited no loss whatsoever, and consequently were eliminated as they could not be used to study loss inference. 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 1.15 times greater than that of the first. Moving further along the stripe, loss rates differed between successive positions typically by up to a factor of 1.02.

*b) Conditional Probabilities.:* Of 70 traces that did exhibit packet loss, 32 had conditional transmission probabilities of 1, reflecting perfect loss correlation just as would occur if the probes had been multicast instead of unicast.

For the remaining 38 traces, we estimate the error involved in the stripe method by comparing conditional and marginal

| | $\beta(1\|2,\ldots,w)/\beta(1\|2,\ldots,w-1)$ | | | | |
|---|---|---|---|---|---|
| | $w=2$ | $w=3$ | $w=4$ | $w=5$ | $w=6$ |
| min. | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| mean | 1.0318 | 1.0017 | 1.0006 | 1.0005 | 1.0003 |
| max. | 1.1812 | 1.0103 | 1.0051 | 1.0031 | 1.0020 |

TABLE II

COALESCENCE OF TRANSMISSION IN NETWORK EXPERIMENTS. RATIOS OF END-TO-END CONDITIONAL TRANSMISSION PROBABILITIES IN STRIPES OF WIDTH 2 TO 6.

transmission probabilities within a stripe. A scatter plot of the conditional vs. marginal probabilities for 2 and 3 packet stripes is shown in Figure 3. (Only 36 points are apparent in the figure due to the occurrence of two pairs of identical loss rates). Higher points represent smaller estimates of 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 $\beta$ are noticeably larger than the marginal transmission probabilities $\alpha$, 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 one signifies perfect correlations. We characterize the error arising from $\beta < 1$ through the ratio $(1-\beta)/(1-\alpha)$ when $\alpha \neq 1$. This represents the proportion of the reported loss rate which is in error due to imperfect correlations. For 2-packet stripes, the median value 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 $0.10$ for 3 packet stripes.

We also verified that $\beta(2|1)/\beta(1|2) \geq 1$ in practice. For the same 38 experiments we computed the ratio using the end-to-end conditional probabilities $\beta(2|1)$ and $\beta(1|2)$. The ratio was one in 19 experiments and overall no greater than 1.0052 in 90% of the traces; in seven instances, it was even smaller than one (but always larger than 0.999); the maximum value was 1.83 (corresponding to the trace with a loss rate of 14%). The fact that the ratio was very close to one can be justified in terms of the traces exhibiting small loss probabilities (26 experiments had loss rates smaller than 1%) for which we expect $\beta \approx 1$ in any case. For a finer comparison of the two conditional probabilities, we also computed the ratio $(1-\beta(2|1))/(1-\alpha)$, $\alpha \neq 1$. The median of this ratio was 0.2, about 66% larger than that due to $\beta(1|2)$. Despite being very similar in these experiments, the likely impact on estimator accuracy of the two conditional probabilities differs substantially.

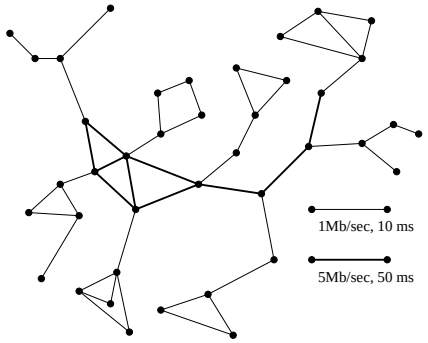


Fig. 4. FIRST TOPOLOGY USED IN SIMULATIONS, COMPRISING 39 NODES

*c) Coalescence.:* We calculated the end-to-end conditional transmission probabilities $\beta(1|2,3,\ldots w)$ for stripes of width $w = 1,\ldots,6$. (When $w = 1$ this just denotes the marginal probability $\alpha(1)$). A necessary condition for coalescence is that $\beta(1|2,\ldots,w)/\beta(1|2,\ldots,w-1) \geq 1$ for all $w$. 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 II: the minimum, mean, and maximum for each $w$ over the 19 experiments. The ratios are largest for $w = 2$, falling off close to one 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 show unicast-based inference to be promising. First, the stripes exhibited perfect loss correlation in nearly half of the traces where there was any loss. If this property were to hold 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 higher than the marginal probabilities, slightly more for the three packet stripe than the two 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 bias can be compensated for by using wider stripes, although the incremental benefit grew smaller for longer stripes. Concluding, we have described a method that can be used to determine, via end to end measurements, whether packet loss correlation within stripes is sufficiently strong (and, in particular, coalescent) for the unicast inference method to be accurate. We found that loss correlation was strong in the network under study.

## V. SIMULATION RESULTS

The experiments of Section IV give us confidence that the statistical properties of stripe transmission make stripes suitable as probes for inference. However, the experiments do not corroborate the accuracy of the estimators for real network traffic. Instead, we employ simulation to illustrate the estimator accuracy that would likely be obtained in a real network setting. We used the `ns` simulation environment [25]; 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.

### A. Simulation Methodology

We conducted simulations using two topologies. The first was the 39-node topology of Figure 4. Link speeds and delays characterize low speed/low delay links at a network edge connected by high speed/high delay links in the network interior. Each link buffer accommodates 20 packets. Background traffic came from a mixture of TCP sessions and exponential and Pareto on-off UDP sources with shape parameter 1.5. The on-off sources had mean burst time 0.5s. The simulation ran for 800 seconds, giving ample time for equilibration on the traffic sources. Simulations were performed using two logical multicast trees spanning 7 and 15 receivers.

The second topology was generated using the `gt-itm` topology generator [15]. It comprised 156 nodes arranged as a hierarchical transit-stub network in which 24 stub networks are interconnected via a 12 node transit network. Links between transit nodes have 50Mb/s capacity and propagation delay chosen randomly in the interval [8ms,20ms]. The other links have a 10Mb/s capacity and a delay chosen randomly in the interval [1ms,10ms]. Without randomized link delays, we would open up the chance for synchronization between traffic flows on different end-to-end paths with identical round trip times [12]. This could potentially lead to violation of the independent loss model. The buffer on each link accommodates 100 packets. We selected a 38 receiver multicast tree comprising 62 nodes. The number of hops between the source and a receiver ranges between 5 and 11 with an average of 7.34. Background traffic was similar to that used in the 39-node topology.

Although other choices of topology could be considered (e.g., those generated by BRITE [2]), since the inference method makes no reference to topology, we do not expect the results to be sensitive to topology, except as follows. Firstly, larger topologies require more measurements in order to cover each node as a branch point. Secondly, diversity in larger topologies is expected to reduce loss synchronization between links, and hence improve the quality of inference.

In both topologies, measurement probes comprised 4 packet stripes with a $1\mu$sec interpacket time. The inter-stripe time was 16 msec, cycling through stripes $S(i,j)$ over pairs of distinct receivers $i,j$. The number of cycles $n$ was chosen so that the total number of stripes $m$ sent was the same in all simulations. This enables us to compare performance for the same measurement traffic load. Here, we chose $m = 42,000$.

Since the stripe width is far shorter than the burst length of the on-off sources, we expect loss within a stripe due to congestion arising from bursting of these source to be strongly correlated, as desired. Congestion periods may encompass multiple stripes, leading to dependence of packet loss between

different stripes. But, similarly to the multicast case analyzed in [3], this effect is not expected to alter the limiting value of the loss estimate as the number of stripes grows large, although it can increase estimator variance relative to a model in which losses across different stripes are independent.

What is the load on the network from probing? Assuming a 60 byte packet size as in Section IV-A, the average probe rate at the source would be 15 KB/sec, i.e, about 1% of the slowest link rate, and roughly equivalent to a high quality Voice-over-IP call. Furthermore, the rate reduces away from the source due to branching of the distribution tree. The maximum burst size is 4, i.e., only 4% of the simulated buffer size in the larger topology, and likely an even smaller quantity relative to buffers in deployed equipment. Altogether, we expect probing to have only a small effect on other network traffic.

How do the durations of measurement compare with typical periods of constancy for loss rates? In our experiments, the $42,000$ stripes would take just over 11 minutes to dispatch. In one study on the dynamics of packet loss, 1 minute averaged loss rates over 11 minute intervals were found to be roughly constant (in the operational sense that they didn't move between bands of a few percent width) about 80% of the time; see [32]. But even the absence of such constancy may well not affect the accuracy of the method greatly. Comparison of directly measured and inferred loss rate in the multicast case shows that when the loss rate fluctuates by a few percent over the measurement interval, averaging in the loss inference quite closely reflects the average measured loss rate; see [4].

To compare the estimator performance under 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 rates as determined by instrumentation of the simulation. Link loss rates were computed considering only the first probe in the stripe.

### B. Transmission Probabilities and Coalescence

We first examine the statistical properties of the underlying link loss processes. Marginal and conditional link transmission rates were determined during 100 experiments on the 7 and 15 receiver topologies, and 10 experiments on the 38 receiver topologies. Link loss rates in these three sets of experiments ranged from 0% to 18%, from 0% to 27% and 0% to 2.6% respectively. Scatter plots of conditional vs. marginal transmission probabilities are shown in Figure 5. Conditional probabilities are considerably higher than marginal probabilities, and mostly strictly increasing in stripe width. Note that unlike $\beta(1|23)$ and $\beta(1|234)$, which are always very close to 1, $\beta(1|2)$ falls considerably below 1 as the loss rate increases. This behavior is in agreement with the analysis in Section III where we observed that, among the conditional probabilities, $\beta(1|2)$ decreases fastest at higher loads. For the 38 receiver tree, $\beta(1|234)$ exceeds $\beta(1|23)$ in only a few cases; mostly they are equal. Thus we expect small benefit in accuracy from increasing the stripe width beyond 3 in this topology.

To summarize the conditional probability structure, we computed the ratio $r_k(w) = \beta(1|2,\ldots,w)/\ \beta(1|2,\ldots,w-1)$ for $w = 2, 3, 4$. The statistics are displayed in Table III. The behavior is similar to that observed in the network experiments where the ratio is largest for $w = 2$ and decreases for larger values of $w$. For the 38 receiver topology, the ratios practically equal 1 for $w > 2$. The larger values observed in the 15 receiver tree are due to the larger spread in the conditional probabilities, which correspond to the higher loss rates. In some cases the ratio was smaller than 1: one ratio for $w = 2$, $7\%$ of the ratios for $w = 3$, and $25\%$ of the ratios for $w = 4$; in these cases, though, the ratios were always very close to 1, and the smaller of the two probabilities larger than 0.99. This behavior is expected since the observed ratios will exhibit some statistical variability. Modeling the observed conditional probability ratio $r_k(w)$ as a Gaussian random variable for each $w$, we ask whether the observed mean value $\widehat{r}_k(w)$ is consistent with a population value less than 1. If so, we cannot conclude that the coalescence property holds. Using the sample standard deviation of the observed ratios over 100 independent simulations $\sigma_{r_k(w)}$, we found the test statistic $\widehat{r}_k(w) - z\sigma_{r_k(w)}/\sqrt{98}$ exceeded 1 for all $k \neq 0$, and $w = 2, 3, 4$, where $z$ was the $99^{\text{th}}$ percentile of the standard normal distribution. Hence the observations are consistent with coalescent transmissions, at a 99% confidence level.

### C. Measures of Inference Accuracy

In order to quantify the accuracy of our estimates, we computed, for each logical link, the estimator bias and standard deviation. For each non-root node $k$, denote by $\alpha_k^{(j)}$ and $\widehat{\alpha}_k^{(j)}$ the actual and inferred transmission probability on link $k$ in the $j$-th simulation, for $j = 1,\ldots,N = 100$. For $k \neq 0$, we compute the estimator bias as $b_k := \frac{1}{N}|\sum_{j=1}^{N} \alpha_k^{(j)} - \widehat{\alpha}_k^{(j)}|$ and standard deviation $\sigma_k = \sqrt{\frac{1}{N-1}\left(\sum_{j=1}^{N}\left(\alpha_k^{(j)} - \widehat{\alpha}_k^{(j)}\right)^2 - b_k^2\right)}$. As a robust summary statistic of the typical bias and standard deviation across the different links, we used the two-sided quartile-weighted median (QWM)

$$(Q_{25} + 2Q_{50} + Q_{75})/4, \tag{20}$$

where $Q_p$ denotes the $p^{th}$ quantile of the set of link estimator bias $\{b_k\}_{k\neq 0}$, or standard deviations $\{\sigma_k\}_{k\neq 0}$.

### D. Accuracy and Probing Strategy

For the 7 receiver tree, we compare two probing strategies. In the *exhaustive striping strategy* we run a complete set of measurements down all embedded two-receiver subtrees. In

| $w=2$ | $w=3$ | $w=4$ |
|---|---|---|
| 0.9999 | 0.9994 | 0.9983 |
| 1.0379 | 1.0043 | 1.0006 |
| 1.4725 | 1.0573 | 1.0189 |

| $w=2$ | $w=3$ | $w=4$ |
|---|---|---|
| 1.0000 | 1.0000 | 1.0000 |
| 1.0034 | 1.0001 | 1.0000 |
| 1.0427 | 1.0043 | 1.0022 |

TABLE III

COALESCENCE OF TRANSMISSION IN SIMULATIONS. MINIMUM, MEAN AND MAXIMUM OF RATIO $\beta(1|2,\ldots,w)/\beta(1|2,\ldots,w-1)$ ACROSS ALL LINKS AND SIMULATIONS: (LEFT) 15 RECEIVER TREE; (RIGHT) 38 RECEIVER TREE.

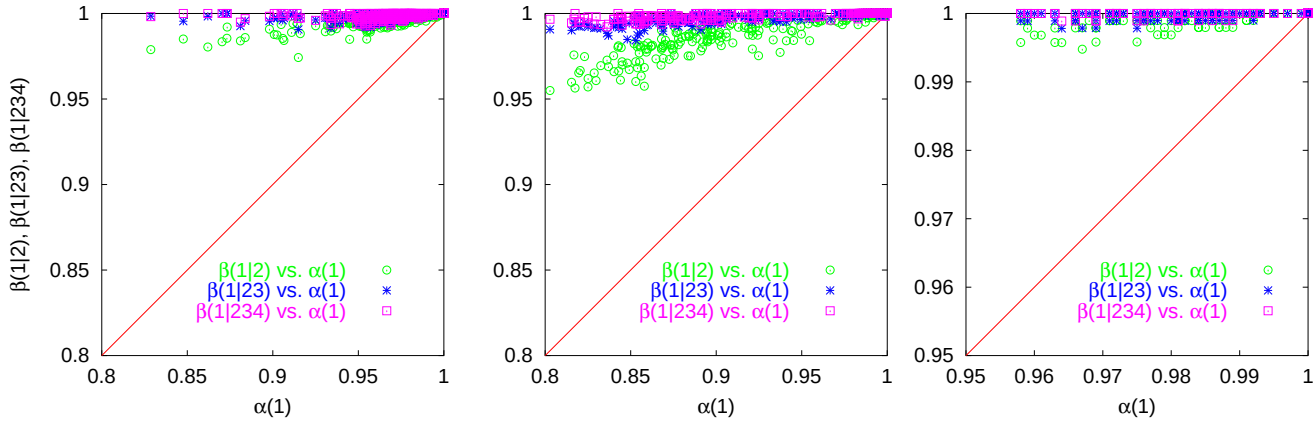


Fig. 5. CONDITIONAL TRANSMISSION PROBABILITIES IN SIMULATIONS. Scatter-plot of conditional vs. marginal link transmission probabilities for 2, 3 and 4 packet stripes: (left) 7 receiver tree; (middle) 15 receiver tree; (right) 38 receiver tree. Conditional probabilities increase with stripe width.

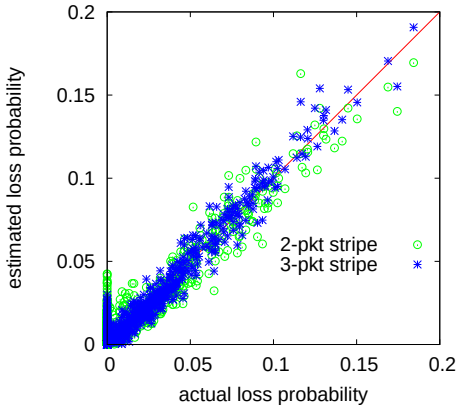


Fig. 6. 7 RECEIVER LOGICAL TREE; EXHAUSTIVE STRIPING STRATEGY. INFERRED VS. ACTUAL LINK LOSS RATES. 3 and 2 packet substripes.

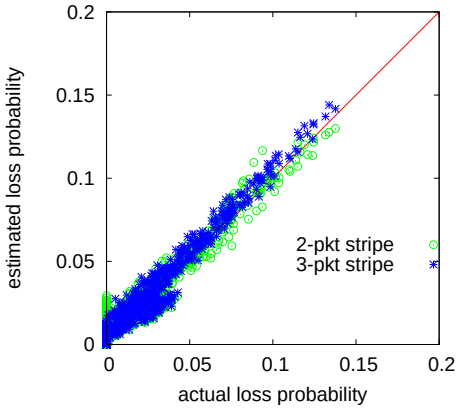


Fig. 7. 7 RECEIVER LOGICAL TREE; MINIMAL STRIPING STRATEGY. INFERRED VS. ACTUAL LINK LOSS RATES. 3 and 2 packet substripes.

this tree there are 21 such subtrees, and so we used all 42 stripes $S(i,j)$ over (ordered) pairs of disjoint receivers. In accordance with the choice of $m = 42,000$ total stripes, each stripe was transmitted 1,000 times. In the *minimal striping strategy*, we select (at random) one subtree through each branch point. In the example there are 5 subtrees and hence 10 ordered pairs of receivers; each stripe was transmitted 4,200 times. We want to determine the trade-off in accuracy between employing more subtrees (in the exhaustive strategy) and more stripes per subtree (in the minimal strategy) for the same number of total stripes.

For the different sets of experiments, we display scatter plots of inferred vs. actual loss probabilities for 2 and 3 packet stripes in Figures 6 for the exhaustive strategy, and Figure 7 for the minimal strategy. From the figures we observe that accuracy increases with wider packet stripes as exhibited by the clustering about the line $y = x$. Accuracy is apparently worse when the actual link loss probability is zero. However, this is a visual effect arising from the large proportion of points (about to 60% of the total) for which the actual loss probability was zero. The standard deviation of the estimates corresponding to zero and non zero loss are actually very close.

| | stripe width | | |
|---|---|---|---|
| | 2 | 3 | 4 |
| bias | 0.58% | 0.35% | 0.31% |
| s.d. | 0.66% | 0.53% | 0.53% |

| | stripe width | | |
|---|---|---|---|
| | 2 | 3 | 4 |
| bias | 0.68% | 0.47% | 0.41% |
| s.d. | 0.47% | 0.29% | 0.29% |

TABLE IV

ESTIMATION ERROR IN SIMULATIONS AS FUNCTION OF STRIPE WIDTH. 7 RECEIVER TREE QWM OF ESTIMATOR BIAS AND STANDARD DEVIATION: (LEFT) EXHAUSTIVE STRIPING; (RIGHT) MINIMAL STRIPING.

Comparing the two striping strategies it appears that points are more tightly clustered about the line $y = x$ for the minimal strategy (Figure 7) than for the exhaustive strategy (Figure 6). To quantify this, we display bias and standard deviation of the QWM in Table IV. Increasing the stripe width reduces bias for both strategies, although most of the benefit is already obtained using a 3 packet stripe. This is not surprising since the largest increase in conditional probabilities occurs when stripe width is increased from 2 to 3. We also performed experiments with 4 stripes and found no further reduction in standard deviation upon increasing the stripe width from 3 to 4. Returning to the comparison of the probing strategies, we see that for larger stripe widths, the standard deviation for the minimal strategy is roughly half that of the exhaustive strategy, although the bias is a little larger. In each strategy the typical absolute error in

| | stripe width | | |
|---|---|---|---|
| | 2 | 3 | 4 |
| bias | 1.51% | 0.39% | 0.26% |
| s.d. | 1.33% | 1.16% | 1.16% |

| stripe width | | |
|---|---|---|
| 2 | 3 | 4 |
| 0.08% | 0.07% | 0.06% |
| 0.18% | 0.17% | 0.17% |

TABLE V

ESTIMATION ERROR IN SIMULATIONS AS FUNCTION OF STRIPE WIDTH. QWM OF ESTIMATOR BIAS AND STANDARD DEVIATION. (LEFT) 15 RECEIVER TREE; (RIGHT) 38 RECEIVER TREE.

loss rate estimation is less than 1%.

### *E. Larger Topologies*

In Table V we display bias and standard deviation of the QWM for simulations on the 15 receiver and 38 receiver topologies; these simulations used the minimal striping strategy. Compared with the 7 receiver tree, the bias in the 15 receiver tree is noticeably larger for $w = 2$ but roughly similar to that observed in the smaller tree otherwise. We verified that the higher value for $w = 2$ is due to larger estimator bias for leaf links. This can be explained by observing that since the bias for receiver links depends on the conditional probability along the entire path from the probe source (see the second relation in (14)), even small departures from unity can result in large bias as the depth of the tree increases. This effect is not noticeable for larger stripe widths since $\beta(1|2,3)$ and $\beta(1|234)$ are, most of the time, equal to one, thus reducing the effect of topology size on the estimator bias of receiver links. The standard deviation was higher for the 15 receiver topology than the 7 receiver topology as a consequence of both the larger size and fewer stripes per subtree.

The smaller underlying loss rates in the 38 receiver tree make it difficult to draw a direct comparison with the smaller trees concerning accuracy. Nevertheless, we find conformance with the pattern that increasing the stripe width from 2 to 3 noticeably reduces bias and standard deviation, but further stripe lengthening achieves little or no further gain in accuracy. Since the conditional probabilities are all very close to 1, here we obtain good accuracy for any stripe length. In distinction with the experiments on smaller topologies, we found that estimation of zero loss rate was noticeably less accurate than that of non-zero loss rates. (This was determined by analysis of the corresponding estimator standard deviations for the two types of link). The large errors in estimation of zero loss rates can be explained by the fact that zero loss occurred mostly at receiver links. This, given the larger variance that we expect for receiver link estimates (as discussed in Section II-E), especially for larger topologies, accounts for the larger variability that we observed.

## VI. 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 multicast packet following 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 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 other packets from the stripe have been successfully transmitted. We proved that coalescence is satisfied for stripes traversing a M/M/1/K queue.

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.

Direct assessment of the method requires corroborative measurements in the network interior. This entails taking measurements on paths over which probe traffic flows; then comparing actual and 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 bias in these experiments was in the worst case about 1.5% in the reported loss rate for the 2-packet stripe, falling to 0.3% with a 4-packet stripe. We believe the accuracy is sufficient to identify the worst performing links down to loss rates of some fraction of 1% in most cases. Most of the benefit in accuracy was obtained using 3 packet stripes; the marginal benefit using 4 packet stripes was relatively small, especially in larger topologies.

In this paper we concentrated on estimation of link probabilities for the first packet of a stripe. Due to heterogeneity of loss along the stripe, such estimates may not be representative of all packets. The present method could be extended to estimate link probabilities for packet in positions other than the first.

Finally, other multicast-based estimators–namely those for delay distributions [23], for delay variances [11], and logical multicast topology [10]–have the potential to be adapted in the same manner as was done for loss estimators in this paper.

*Acknowledgment*

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

## VII. PROOFS OF THEOREMS

**Proof of Theorem 1:** $\mathsf{E}Z_{D\cup D'} = \beta_{\mathsf{c}}(D|D')\alpha_{\mathsf{c}}(D')\,\alpha_{\mathsf{l}}(D)\alpha_{\mathsf{r}}(D')$ while $\mathsf{E}Z_{D'} = \alpha_{\mathsf{c}}(D')\alpha_{\mathsf{r}}(D')$. Hence $\mathsf{E}Z_{D\cup D'}/\mathsf{E}Z_{D'} = \beta_{\mathsf{c}}(D|D')\alpha_{\mathsf{l}}(D) \geq \beta_{\mathsf{c}}(D|D'')\alpha_{\mathsf{l}}(D) = \mathsf{E}Z_{D\cup D''}/\mathsf{E}Z_{D''}$. ■

**Proof of Theorem 2:** Since the random variables $\widetilde{Z}_D$ are the average of i.i.d. random variables $Z_D^{(i)}$, for any $D \subset D_0$, any node $k \neq 0$ and stripe $S(\mathsf{r}_{i_1}, \mathsf{r}_{i_2})$, $(\mathsf{r}_{i_1}, \mathsf{r}_{i_2}) \in Q(k)$, then by Central Limit Theorem $\sqrt{n} \cdot (\widetilde{Z} - Z)$, where $\widetilde{Z} = \{\widetilde{Z}_D\}_{D\subset D_0,(\mathsf{r}_{i_1},\mathsf{r}_{i_2})\in Q(k),k\neq 0}$ and $Z = \{Z_D\}_{D\subset D_0,(\mathsf{r}_{i_1},\mathsf{r}_{i_2})\in Q(k),k\neq 0}$, converges in distribution

to a multivariate Gaussian random variable as $n \to \infty$. Since $\widehat{\alpha}_k$ is a differentiable function $\mathcal{F}_k$ of $\widetilde{Z}$, $\mathcal{F}_k(\widetilde{Z}) = \frac{1}{M(k)} \sum_{(i,j)\in M(k)} \frac{\widetilde{Z}_{D_1}\widetilde{Z}_{D_2}}{\widetilde{Z}_{D_1\cup D_2}}$, the Delta method (see Chapter 7 of [28]) ensures the convergence of $\sqrt{n}(\widehat{\alpha}_k - \mathcal{F}_k(\mathsf{E}Z))$ to a multivariate Gaussian random variable with mean 0. Theorem 2 follows from the stated convergence and because for $k \notin R$, $\mathcal{F}_k(\mathsf{E}Z) = (A_k/B_k)/(A_{f(k)}/B_{f(k)}) = \alpha_k/\beta_k$, and that for $k \in R$, $\mathcal{F}_k(\mathsf{E}Z) = A_k/(A_{f(k)}/B_{f(k)}) = \alpha_k B_{f(k)}$. ■

**Proof of Theorem 3:** The proof is based on the result below which establishes the stochastic order relations among the number of packets in the queue seen by the different probes upon arrival. For random vectors $X$ and $Y$, we say that $X$ is smaller than $Y$ in *stochastic order* (denoted $X \leq_{st} Y$), if $\mathsf{E}[h(X)] \leq \mathsf{E}[h(Y)]$ for any function $h$, nondecreasing in each argument, for which expectation exists. In case $X, Y \in \mathbb{R}$, this is equivalent to the condition $\mathsf{P}[X \leq x] \geq \mathsf{P}[Y \leq x]\ \forall x$.

Let $N$ denote the steady state number of packets in the M/M/1/K queue fed by background traffic only. Let $N_d^S$ denote the number of users in the system seen by the $d$-th probe upon arrival given the set of probes $S \subseteq D_0$ are not lost and let $N^S = \{N_1^S, \ldots, N_{d_0}^S\}$. The following holds.

*Theorem 4:* In a M/M/1/K queue,

(i) $N_1 \leq_{st} N_2 \leq_{st} \ldots \leq_{st} N_{d_0}$;
(ii) $N_1^2 \leq_{st} N_2^1$;
(iii) for any $R \subset S \subseteq D_0$, $N^R \leq_{st} N^S$.

Theorem 3 is then an immediate consequence of Theorem 4. To prove the coalescence property, for any disjoint $D, D' \subset D_0$ we can write $1 - \beta(D|D') = \mathsf{E}[\mathbf{1}_{\{\vee_{d\in D} N_d^{D'} > K\}}]$, where $\mathbf{1}_{\{.\}}$ denotes the indicator function. Then, Theorem 4(iii) implies that for $D'' \subset D' \subseteq D_0$, $N^{D'} \leq_{st} N^{D''}$ which, coupled with the fact that $\mathbf{1}_{\{\vee_{d\in D} N^{D'} > K\}}$ is nonincreasing in each argument, yields $1 - \beta(D|D') \leq 1 - \beta(D|D'')$, *i.e.*, $\beta(D|D') \geq \beta(D|D'')$. Transmission is thus coalescent. Theorem 3 parts (i) and (ii) follow from (i) and (ii) above since $\alpha(d) = \mathsf{P}[N_d \leq K]$, $d \in D_0$, and that $\beta(2|1) = \mathsf{P}[N_2^1 \leq K]$.

**Proof of Theorem 4:** The proof proceeds through sample path arguments. In the following, let $Q(t)$ denote the queue length at time $t$. If an arrival occurs at time $t$, $Q(t^-)$ will denote the queue length just before the arrival, *i.e.*, the queue length seen by the arriving packet. Last, let $\{t_d\}_{d=1}^{d_0}$ be the arrival times of the probes. By definition, $N_i = Q(t_i^-)$.

(i). We first show that $N_1 \leq_{st} N_2$. Recall that we assume the first probe finds the queue in steady state, *i.e.* that $Q(t_1^-) =_{st} N_1 =_{st} N$. Then, it immediately follows that

$$\mathsf{P}[N \leq m] \geq \mathsf{P}[Q(t_1) \leq m] \quad m = 0, \ldots K+1. \tag{21}$$

We now consider a second, benchmark, M/M/1/K queue with only background traffic arrivals. Let $Q'(t)$ denote the queue length at time $t$ and assume that it has reached steady state by $t = t_1$. We now couple the two systems, the first where the first probe arrives to the queue at $t = t_1$, the second the benchmark M/M/1/K queue. The inequality in (21) allows us to couple the systems so that $Q(t_1) \geq Q'(t_1)$. We now create a sequence of event times from a Poisson process with rate $\lambda_b + \mu$ from $t = t_1$ up until $t = t_2$. Let these times be $\tau_1 < \tau_2 < \cdots < \tau_k$. Associate with each of these either an arrival event or service completion event with probabilities $\lambda_b/(\lambda_b+\mu)$ and $\mu/(\lambda_b+\mu)$ respectively. The systems now behave as follows. If there is an arrival at time $\tau_i$, then

$$\begin{aligned} Q(\tau_i) &= \min(Q(\tau_{i-1}) + 1, K) \\ Q'(\tau_i) &= \min(Q'(\tau_{i-1}) + 1, K) \end{aligned}$$

Similarly, if there is a service completion, then

$$\begin{aligned} Q(\tau_i) &= \max(Q(\tau_{i-1}) - 1, 0) \\ Q'(\tau_i) &= \max(Q'(\tau_{i-1}) - 1, 0) \end{aligned}$$

A simple induction argument on the event times allows us to show that $Q(t) \geq Q'(t)$ for $t_1 \leq t < t_2$. (Note that the exponential time assumption is needed to ensure that arrival and service completions can be coupled to the Poisson event process in the manner indicated). Thus, for $t = t_2$, $N_2 = Q(t_2^-) \geq N(t_2^-)$. Removing the conditioning on the initial queue lengths, arrivals, and service completions yields $N(t_2^-) \leq_{st} N_2$. Since the second queue is a M/M/1/K queue in steady state, $N(t_2^-) =_{st} N(t_1) =_{st} N$. Therefore, $N_1 =_{st} N \leq_{st} N_2$. Similar arguments can be used to establish the remaining stochastic inequalities in (i).

(ii). We establish (ii) by first showing that $N \leq_{st} N_2^1$ and then that $N_1^2 \leq_{st} N$. It is easy to verify that

$$\mathsf{P}[N \leq m] \geq \mathsf{P}[Q(t_1) \leq m | Q(t_1^-) \leq K],\ m \leq K+1. \tag{22}$$

We now couple the two systems, the first where the first probe successfully made it into the queue at $t = t_1$, the second the benchmark M/M/1/K queue as above. The inequality in (22) allows us to couple the systems so that $Q(t_1) \geq Q'(t_1)$. Using the same arguments as above, it then $N \leq_{st} N_2^1$.

Consider now the relation $N_1^2 \leq_{st} N$. We make use of the fact that $N_2^2 \leq_{st} N_2$ and that the M/M/1/K queue is modeled by a time-reversible Markov chain. Consider our original system with probes arriving at exponentially distributed intervals starting at $t_1$ and ending at $t_{d_0}$. The system behaves as an M/M/1/K queue in the interval $[t_1, t_{d_0}]$. We focus on the time reversed behavior during the interval $[t_1, t_2]$. We consider two systems, one where the second probe is known to have been accepted ($N_2^2 < K$) and the other where no information is known about the second probe. Now, $\mathsf{P}[N_2^2 \leq m] = \mathsf{P}[N_2 \leq m | N_2 \leq K]$. Therefore, $N_2^2 \leq_{st} N_2$. We now couple the queue lengths of the two systems so that $Q(t_2^-) \leq Q'(t_2^-)$, where $\{Q(t)\}$ and $\{Q'(t)\}$ are the queue length processes of these systems. We then couple the time-reversed systems where departures (resp. arrivals) within the original systems during $[t_1, t_2]$ are coupled to arrivals (resp. departures) within the reversed systems. Using reverse induction on $\tau_1, \ldots, \tau_k$, and the arguments used to establish (i), we can conclude that $Q(t^-1) \leq Q'(t_1^-)$. Coupled with the fact that $Q(t_1^-) =_{st} N$, the queue length of an M/M/1/K queue without probes, we conclude that $N_2^1 \leq_{st} N_1 = N$.

(iii). It suffices to show the inequality for $R$, $S$ such that $S = R \cup \{i\}$. First, $\mathsf{P}[N_i^S \leq m] = \mathsf{P}[N_i^R \leq m | N_i^R \leq K]$; thus, $N_i^S \leq_{st} N_i^R$. To establish the result, we couple two M/M/1/K queues at $t = t_1$ where the probes in $S$ make it into the first queue, and probes in $R$ make it into the second queue. Since

$N_i^S \leq_{st} N_i^R$, we can couple the systems so that their queue lengths, denoted by $Q^S(t)$ and $Q^R(t)$, obey $Q^S(t_i) \leq Q^R(t_i)$.

It remains to show what happens for $t > t_i$ and $t < t_i$. We now couple arrivals and service completions (conditioned on successful arrivals of the probes in $S \cap \{i+1, \ldots, d_0\}$) for the two systems until the first time $s$ that $Q^S(s) = Q^R(s)$. At that point in time, the Markov property allows us to couple the two systems so that $Q^S(t) = Q^R(t)$ for $t > s$ thus yielding $Q^S(t) \leq Q^R(t)$ for $t \geq t_i$. Again, the facts that *(i)* the system during the interval $[t_1, t_{d_0}]$ behaves like an M/M/1/K queue and *(ii)* the M/M/1/K queue is modeled by a time-reversible Markov chain, allow us to use a similar argument to show that $Q^S(t) \leq Q^R(t)$ for $t < t_i$. Removal of the conditioning on the arrivals and service completions yields $N^S \leq_{st} N^R$. ■

## REFERENCES


[1] A. Adams, T. Bu, R. Cáceres, 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] BRITE: Boston university Representative Internet Topology gEnerator. See http://www.cs.bu.edu/brite/

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

[4] R. Caceres, N.G. Duffield, S.B. Moon, D. Towsley, "Inferring link-level performance from end-to-end multicast measurements", Global Internet, Rio de Janiero, 1999.

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

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

[7] M.J. Coates and R. Nowak, "Network Delay Distribution Inference from End-to-end Unicast Measurement," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2001.

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

[9] N.G. Duffield, J. Horowitz, D. Towsley, W. Wei, T. Friedman, "Multicast-based loss inference with missing data", IEEE Journal on Selected Areas in Communications, 2002.

[10] N.G. Duffield, J. Horowitz, F. Lo Presti, D. Towsley, Multicast Topology Inference from Measured End-to-end Loss, IEEE Transactions in Information Theory, 48:26–45, 2002.

[11] 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.

[12] S. Floyd, V. Jacobson, "On traffic phase effects in packet-switched gateways," Internetworking: Research and Experience, 3:115–156, 1992.

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

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

[15] GT-ITM Georgia Tech Internetwork Topology Models. For more information see http://www.cc.gatech.edu/fac/Ellen.Zegura/graphs.html

[16] K. Harfoush, A. Bestavros, J. Byers. Robust identification of shared losses using end-to-end unicast probes. Proc. ICNP 2000, Nov. 2000.

[17] G. He, J.C. Hou. On exploiting long range dependence of network traffic in measuring cross traffic on an end-to-end basis. INFOCOM'03, April 2003.

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

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

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

[21] K. Lai, M. Baker,"Measuring link bandwidths using a deterministic model of packet delay," *Proc. SIGCOMM 2000*, Sweden, August 2000.

[22] 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.

[23] F. Lo Presti, N.G. Duffield, J. Horowitz, D. Towsley, "Multicast-based inference of network-internal delay distributions", IEEE/ACM Transactions on Networking, 10:761–775, 2002.

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

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

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

[27] 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.

[28] M.J. Schervish, "Theory of Statistics", Springer, New York, 1995.

[29] V. Ribeiro, M. Coates, R. Riedi, S. Sarvotham, B. Hendricks, R. Baraniuk. Multifractal cross-traffic estimation. ITC Specialist Seminar on IP Traffic Measurement, Modeling, and Management, Sept. 2000.

[30] D. Rubenstein, J. Kurose, D. Towsley. Detecting shared congestion of flows via end-to-end measurement. ACM SIGMETRICS, June 2000.

[31] Y. Tsang, M.J. Coates and R. Nowak, "Passive Network Tomography using EM Algorithms," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2001.

[32] Y. Zhang, N.G. Duffield, V. Paxson, S. Shenker, "On the Constancy of Internet Path Properties", ACM SIGCOMM Internet Measurement Workshop 2001, San Francisco, CA, November 1-2, 2001.