Multi-Hop Probing Asymptotics in Available Bandwidth Estimation: Stochastic Analysis

Xiliang Liu
City University of New York
xliu@gc.cuny.edu

Kaliappa Ravindran
City College of New York
ravi@cs.ccny.cuny.edu

Dmitri Loguinov
Texas A&M University
dmitri@cs.tamu.edu

Abstract

This paper analyzes the asymptotic behavior of packet-train probing over a multi-hop network path $\mathcal{P}$ carrying arbitrarily routed bursty cross-traffic flows. We examine the statistical mean of the packet-train output dispersions and its relationship to the input dispersion. We call this relationship the *response curve* of path $\mathcal{P}$. We show that the real response curve $\mathcal{Z}$ is *tightly* lower-bounded by its *multi-hop fluid counterpart* $\mathcal{F}$, obtained when every cross-traffic flow on $\mathcal{P}$ is hypothetically replaced with a constant-rate fluid flow of the same average intensity and routing pattern. The real curve $\mathcal{Z}$ asymptotically approaches its fluid counterpart $\mathcal{F}$ as probing packet size or packet train length increases. Most existing measurement techniques are based upon the single-hop fluid curve $\mathcal{S}$ associated with the bottleneck link in $\mathcal{P}$. We note that the curve $\mathcal{S}$ coincides with $\mathcal{F}$ in a certain large-dispersion input range, but falls below $\mathcal{F}$ in the remaining small-dispersion input ranges. As an implication of these findings, we show that bursty cross-traffic in multi-hop paths causes negative bias (asymptotic underestimation) to most existing techniques. This bias can be mitigated by reducing the deviation of $\mathcal{Z}$ from $\mathcal{S}$ using large packet size or long packet-trains. However, the bias is not completely removable for the techniques that use the portion of $\mathcal{S}$ that falls below $\mathcal{F}$.

1 Introduction

End-to-end estimation of the spare capacity along a network path using packet-train probing has recently become an important Internet measurement research area. Several measurement techniques such as TOPP [14], Pathload [6], IGI/PTR [5], Pathchirp [16], and Spruce [17] have been developed. Most of the current proposals use a single-hop path with constant-rate fluid cross-traffic to justify their methods. The behavior and performance of these techniques in a multi-hop path with general bursty cross-traffic is limited to experimental evaluations. Recent work [9] initiated the effort of developing an analytical foundation for bandwidth measurement techniques. Such a foundation is important in that it helps achieve a clear understanding of both the validity and the inadequacy of current techniques and provides a guideline to improve them. However, the analysis in [9] is restricted to single-hop paths. There is still a void to fill in understanding packet-train bandwidth estimation over a multi-hop network path.

Recall that the available bandwidth of a network hop is its residual capacity after transmitting cross-traffic within a certain time interval. This metric varies over time as well as a wide range of observation time intervals. However, in this paper, we explicitly target the measurement of a *long-term average* available bandwidth, which is a stable metric independent of observation time instances and observation time intervals [9]. Consider an N-hop network path $\mathcal{P} = (L_1, L_2, \ldots, L_N)$, where the capacity of link L_i is denoted by C_i and the long-term average of the cross-traffic arrival rate at L_i is given by λ_i, which is assumed to be less than C_i. The hop available bandwidth of L_i is $A_i = C_i - \lambda_i$. The path available bandwidth $A_{\mathcal{P}}$ is given by

$$A_{\mathcal{P}} = \min_{1 \le i \le N} (C_i - \lambda_i). \quad (1)$$

The hop L_b, which carries the minimum available bandwidth, is called the *tight* link or the bottleneck link[1]. That is,

$$b = \arg \min_{1 \le i \le N} (C_i - \lambda_i). \quad (2)$$

The main idea of packet-train bandwidth estimation is to infer $A_{\mathcal{P}}$ from the relationship between the inter-packet dispersions of the output packet-trains and those of the input packet-trains. Due to the complexity of this relationship in arbitrary network paths with bursty cross-traffic flows, previous work simplifies the analysis using a single-hop path with fluid[2] cross-traffic, while making the following two assumptions without formal justification: first, cross-traffic burstiness only causes measurement variability that can be smoothed out by averaging multiple probing sam-

ples and second, non-bottleneck links have negligible impact on the proposed techniques.

The validity of the first assumption is partially addressed in [9], where the authors use a single-hop path with bursty cross-traffic to derive the statistical mean of the packet-train output dispersions as a function of the input probing dispersion, referred to as the single-hop response curve. Their analysis shows that besides measurement variability, cross-traffic burstiness can also cause *measurement bias* to the techniques that are based on fluid analysis. This measurement bias *cannot* be reduced even when an infinite number of probing samples are used, but can be mitigated using long packet-trains and/or large probing packet size.

This paper addresses further the two assumptions that current techniques are based on. To this end, we extend the asymptotic analysis in [9] to arbitrary network paths and uncover the nature of the measurement bias caused by bursty cross-traffic flows in a *multi-hop* network path. This problem is significantly different from previous single-hop analysis due to the following reasons. First, unlike single-hop measurements, where the input packet-trains have deterministic and equal inter-packet separation formed by the probing source, the input packet-trains at any hop (except the first one) along a multi-link path are output from the previous hop and have random structure. Second and more importantly, the multi-hop probing asymptotics are strongly related to the routing pattern of cross-traffic flows. This issue never arises in a single-hop path and it has received little attention in prior investigation. However, as we show in this paper, it is one of the most significant factors that affect the accuracy of bandwidth measurement in multi-hop paths.

To characterize packet-train bandwidth estimation in its most general settings, we derive the probing response curve $\mathcal{Z}$ of a multi-hop path $\mathcal{P}$ assuming arbitrarily routed bursty cross-traffic flows. We compare $\mathcal{Z}$ with its multi-hop fluid counterpart $\mathcal{F}$, which is a response curve obtained when every cross-traffic flow in $\mathcal{P}$ is hypothetically replaced with a fluid flow of the same average intensity and routing pattern. We show, under an ergodic stationarity assumption for each cross-traffic flow, that the real curve $\mathcal{Z}$ is tightly lower bounded by its fluid counterpart $\mathcal{F}$ and that the curve $\mathcal{Z}$ asymptotically approaches its fluid bound $\mathcal{F}$ in the entire input range as probing packet size or packet-train length increases.

Most of the existing techniques are based on the single-hop fluid response curve $\mathcal{S}$ associated with the bottleneck link in $\mathcal{P}$. Therefore, any deviation of the real curve $\mathcal{Z}$ from the single-hop curve $\mathcal{S}$ can potentially cause measurement bias in bandwidth estimation. Note that the deviation $\mathcal{Z}-\mathcal{S}$ can be decomposed as

$$\mathcal{Z} - \mathcal{S} = (\mathcal{Z} - \mathcal{F}) + (\mathcal{F} - \mathcal{S}). \quad (3)$$

The first term $\mathcal{Z} - \mathcal{F}$ is always positive and causes asymptotic underestimation of $A_{\mathcal{P}}$ for most of the existing techniques. This deviation term and its resulting measurement bias are "elastic" in the sense that they can be reduced to a negligible level using packet-trains of sufficient length[3]. For the second deviation term $\mathcal{F} - \mathcal{S}$, we note that both $\mathcal{S}$ and $\mathcal{F}$ are piece-wise linear curves. The first two linear segments in $\mathcal{F}$ associated with large input dispersions coincide with $\mathcal{S}$ (i.e., $\mathcal{F} - \mathcal{S} = 0$). The rest of the linear segments in $\mathcal{F}$ associated with small input dispersions appear above $\mathcal{S}$ (i.e., $\mathcal{F} - \mathcal{S} > 0$). The amount of deviation and the additional negative measurement bias it causes are dependent on the routing patterns of cross-traffic flows, and are maximized when every flow traverses only one hop along the path (which is often called *one-hop persistent* cross-traffic routing [4]). Furthermore, the curve deviation $\mathcal{F}-\mathcal{S}$ is "non-elastic" and stays constant with respect to probing packet size and packet-train length at any given input rate. Therefore, the measurement bias it causes cannot be overcome by adjusting the input packet-train parameters.

Among current measurement techniques, pathload and PTR operate in the input probing range where $\mathcal{F}$ coincides with $\mathcal{S}$, and consequently are only subject to the measurement bias caused by the first deviation term $\mathcal{Z} - \mathcal{F}$. Spruce may use the probing range where $\mathcal{F} - \mathcal{S} > 0$. Hence it is subject to both elastic and non-elastic negative measurement biases. The amount of bias can be substantially more than the actual available bandwidth in certain common scenarios, leading to negative results by the measurement algorithm and a final estimate of zero by the tool.

The rest of the paper is organized as follows. Section 2 derives the multi-hop response curve $\mathcal{F}$ assuming arbitrarily routed fluid cross-traffic flows and examines the deviation term $\mathcal{F} - \mathcal{S}$. In Section 3 and 4, we derive the real response curve $\mathcal{Z}$ of a multi-hop path and show its relationship to its fluid counterpart $\mathcal{F}$. We provide practical evidence for our theoretical results using testbed experiments and real Internet measurements in Section 5. We examine the impact of these results on existing techniques in Section 6 and summarize related work in Section 7. Finally, we briefly discuss future work and conclude in Section 8.

Due to limited space, most of the proofs in this paper are omitted, and we refer interested readers to [10] for more technical details.

2 Multi-Hop Fluid Analysis

It is important to first thoroughly understand the response curve $\mathcal{F}$ of a network path carrying fluid cross-traffic flows, since as we show later, the fluid curve $\mathcal{F}$ is an *approachable* bound of the real response curve $\mathcal{Z}$. Initial investigation of the fluid curves is due to Melandar *et al.* [13] and Dovrolis *et al.* [3]. However, prior work only considers two special cross-traffic routing cases (one-hop persistent routing and path persistent routing). In this section, we formulate

and solve the problem for arbitrary cross-traffic routing patterns, based on which, we discuss several important properties of the fluid response curves that allow us to obtain the path available bandwidth information.

2.1 Formulating A Multi-Hop Path

We first introduce necessary notations to formulate a multi-hop path and the cross-traffic flows that traverse along the path.

An N-hop network path $\mathcal{P} = (L_1, L_2, \ldots, L_N)$ is a sequence of N interconnected *First-Come First-Served (FCFS) store-and-forward* hops. For each forwarding hop L_i in $\mathcal{P}$, we denote its link capacity by C_i, and assume that it has infinite buffer space and a work-conserving queuing discipline. Suppose that there are M fluid cross-traffic flows traversing path $\mathcal{P}$. The rate of flow j is denoted by x_j and the flow rate vector is given by $\mathbf{x} = (x_1, x_2, \ldots, x_M)$.

We impose two routing constraints on cross-traffic flows to simplify the discussion. The first constraint requires every flow to have a different routing pattern. In the case of otherwise, the flows with the same routing pattern should be aggregated into one single flow. The second routing constraint requires every flow to have only one link where it enters the path and also have only one (downstream) link where it exits from the path. In the case of otherwise, the flow is decomposed into several separate flows that meet this routing constraint.

Definition 1 *A flow aggregation is a set of flows, represented by a "selection vector"* $\mathbf{p} = (p_1, p_2, \ldots, p_M)^T$*, where* $p_j = 1$ *if flow* j *belongs to the aggregation and* $p_j = 0$ *if otherwise. We use* $\mathbf{f}_j$ *to represent the selection vector of the aggregation that contains flow* j *alone.*

There are several operations between flow aggregations. First, the common flows to aggregations $\mathbf{p}$ and $\mathbf{q}$ form another aggregation, whose selection vector is given by $\mathbf{p} \odot \mathbf{q}$, where the operator $\odot$ represents "element-wise multiplication." Second, the aggregation that contains the flows in $\mathbf{p}$ but not in $\mathbf{q}$ is given by $\mathbf{p} - \mathbf{p} \odot \mathbf{q}$. Finally, note that the traffic intensity of aggregation $\mathbf{p}$ can be computed from the inner product $\mathbf{xp}$.

We now define several types of flow aggregation frequently used in this paper. First, the traversing flow aggregation at link L_i, denoted by its selection vector $\mathbf{r}_i$, includes all fluid flows that pass through L_i. The $M \times N$ matrix $\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$ becomes the routing matrix of path $\mathcal{P}$. For convenience, we define an auxiliary selection vector $\mathbf{r}_0 = \mathbf{0}$.

The second type of flow aggregation, denoted by $\mathbf{e}_i$, includes all flows entering the path at link L_i, which can be expressed as $\mathbf{e}_i = \mathbf{r}_i - \mathbf{r}_i \odot \mathbf{r}_{i-1}$ given the second routing constraint stated previously. The third type of flow aggregation, which includes flows that enter the path at link L_k and traverse the downstream link L_i, is denoted as $\Gamma_{k,i} = \mathbf{e}_k \odot \mathbf{r}_i$, where $k \leq i$.

The cross-traffic intensity at link L_i is denoted by λ_i. We assume $\lambda_i < C_i$ for $1 \leq i \leq N$. Since none of the links in $\mathcal{P}$ is congested, the arrival rate of flow j at any link it traverses is x_j. Consequently, we have

$$\lambda_i = \mathbf{x}\mathbf{r}_i < C_i, \quad 1 \leq i \leq N. \tag{4}$$

We further define the *path configuration* of $\mathcal{P}$ as the following $2 \times N$ matrix

$$\mathbf{H} = \begin{pmatrix} C_1 & C_2 & \ldots & C_N \\ \lambda_1 & \lambda_2 & \ldots & \lambda_N \end{pmatrix}. \tag{5}$$

The hop available bandwidth of L_i is given by $A_i = C_i - \lambda_i$. We assume that every hop has different available bandwidth, and consequently that the tight link is unique. Sometimes, we also need to refer to the second minimum hop available bandwidth and the associated link, which we denote as $A_{b2} = C_{b2} - \lambda_{b2}$ and L_{b2}, respectively. That is

$$b2 = \arg \min_{1 \leq i \leq N, i \neq b} (C_i - \lambda_i), \tag{6}$$

where b is the index of the tight hop.

2.2 Fluid Response Curves

We now consider a packet-train of input dispersion (i.e., inter-packet spacing) g_I and packet size s that is used to probe path $\mathcal{P}$. We are interested in computing the output dispersion of the packet train and examining its relation to g_I. Such a relation is called the *gap response curve* of path $\mathcal{P}$. It is easy to verify that under fluid conditions, the response curve does not depend on the packet-train length n. Hence, we only consider the case of packet-pair probing. We denote the output dispersion at link L_i as $\gamma_i(g_I, s)$ or γ_i for short, and again for notational convenience we let $\gamma_0 = g_I$. Note that $\gamma_N(g_I, s)$ corresponds to the notation $\mathcal{F}$ we have used previously.

Based on our formulations, the gap response curve of path $\mathcal{P}$ has a recursive representation given below.

Theorem 1 *When a packet-pair with input dispersion* g_I *and packet size* s *is used to probe an* N*-hop fluid path with routing matrix* $\mathbf{R}$ *and flow rate vector* $\mathbf{x}$*, the output dispersion at link* L_i *can be recursively expressed as*

$$\gamma_i = \begin{cases} g_I & i = 0 \\ \max\left(\gamma_{i-1}, \dfrac{s + \Omega_i}{C_i}\right) & i > 0 \end{cases}, \tag{7}$$

where Ω_i *is* [4]

$$\Omega_i = \sum_{k=1}^{i} \Big[\gamma_{k-1}\mathbf{x}\Gamma_{k,i}\Big]. \tag{8}$$

Proof: Assumes that the first probing packet arrives at link L_i at time instance a_1. It gets immediate transmission service and departs at a_1+s/C_i. The second packet arrives at $a_1+\gamma_{i-1}$. The server of L_i needs to transmit $s+\Omega_i$ amount of data before it can serve the second packet. If this is done before time instance $a_1+\gamma_{i-1}$, the second packet also gets immediate service and $\gamma_i=\gamma_{i-1}$. Otherwise, the sever undergoes a busy period between the departure of the two packets, meaning that $\gamma_i=(s+\Omega_i)/C_i$. Therefore, we have

$$\gamma_i = \max\left(\gamma_{i-1}, \frac{s+\Omega_i}{C_i}\right). \quad (9)$$

This completes the proof of the theorem. ■

As a quick sanity check, we verify the compatibility between Theorem 1 and the special one-hop persistent routing case, where every flow that enters the path at link L_i will exit the path at link L_{i+1}. For this routing pattern, we have

$$\Gamma_{k,i} = \begin{cases} \mathbf{0} & i \neq k \\ \mathbf{r}_i & i = k \end{cases}. \quad (10)$$

Therefore, equation (8) can be simplified as

$$\Omega_i = \gamma_{i-1}\mathbf{x}\mathbf{r}_i = \gamma_{i-1}\lambda_i, \quad (11)$$

which agrees with previous results [3], [13].

2.3 Properties of Fluid Response Curves

Theorem 1 leads to several important properties of the fluid response curve $\mathcal{F}$, which we discuss next. These properties tell us how bandwidth information can be extracted from the curve $\mathcal{F}$, and also show the deviation of $\mathcal{F}$, as one should be aware of, from the single-hop fluid curve $\mathcal{S}$ of the tight link.

Property 1 *The output dispersion $\gamma_N(g_I,s)$ is a continuous piece-wise linear function of the input dispersion g_I in the input dispersion range $(0,\infty)$.*

Let $0=\alpha_{K+1}<\alpha_K<\ldots<\alpha_1<\alpha_0=\infty$ be the input dispersion turning points that split the gap response curve to $K+1$ linear segments[5]. Our next result discusses the turning points and linear segments that are of major importance in bandwidth estimation.

Property 2 *The first turning point α_1 corresponds to the path available bandwidth in the sense that $A_{\mathcal{P}}=s/\alpha_1$. The first linear segment in the input dispersion range $(\alpha_1=s/A_{\mathcal{P}},\infty)$ has slope 1 and intercept 0. The second linear segment in the input dispersion range (α_2,α_1) has slope λ_b/C_b and intercept s/C_b, where b is the index of the tight link:*

$$\gamma_N(g_I,s) = \begin{cases} g_I & \alpha_1 \leq g_I \leq \infty \\ \dfrac{g_I\lambda_b+s}{C_b} & \alpha_2 \leq g_I \leq \alpha_1 \end{cases}. \quad (12)$$

These facts are irrespective of the routing matrix.

It helps to find the expression for the turning point α_2, so that we can identify the exact range for the second linear segment. However, unlike α_1, the turning point α_2 is dependent on the routing matrix. In fact, all other turning points are dependent on the routing matrix and can not be computed based on the path configuration matrix alone. Therefore, we only provide a bound for α_2.

Property 3 *For any routing matrix, the term s/α_2 is no less than A_{b2}, which is the second minimum hop available bandwidth of path $\mathcal{P}$.*

The slopes and intercepts for all but the first two linear segments are related to the routing matrix. We skip the derivation of their expressions, but instead provide both a lower bound and an upper bound for the entire response curve.

Property 4 *For a given path configuration matrix, the gap response curve associated with any routing matrix is lower bounded by the single-hop gap response curve of the tight link*

$$\mathcal{S}(g_I,s) = \begin{cases} g_I & g_I > \dfrac{s}{A_{\mathcal{P}}} \\ \dfrac{s+g_I\lambda_b}{C_b} & 0 < g_I < \dfrac{s}{A_{\mathcal{P}}} \end{cases}. \quad (13)$$

It is upper bounded by the gap response curve associated with one-hop persistent routing.

We now make several observations regarding the deviation of $\gamma_N(g_I,s)$ (i.e., $\mathcal{F}$) from $\mathcal{S}(g_I,s)$. Combing (12) and (13), we see that $\gamma_N(g_I,s)-\mathcal{S}(g_I,s)=0$ when $g_I\geq\alpha_2$. That is, the first two linear segments on $\mathcal{F}$ coincide with $\mathcal{S}$. When $g_I<\alpha_2$, Property 4 implies that the deviation $\gamma_N(g_I,s)-\mathcal{S}(g_I,s)$ is positive. The exact value depends on cross-traffic routing and it is maximized in one-hop persistent routing for any given path configuration matrix.

Also note that there are three pieces of path information that we can extract from the gap response curve $\mathcal{F}$ without knowing the routing matrix. By locating the first turning point α_1, we can compute the path available bandwidth. From the second linear segment, we can obtain the tight link capacity and cross-traffic intensity (and consequently, the bottleneck link utilization) information. Other parts of the response curve $\mathcal{F}$ are less readily usable due to their dependence on cross-traffic routing.

2.4 Rate Response Curves

To extract bandwidth information from the output dispersion γ_N, it is often more helpful to look at the *rate* response curve, i.e., the functional relation between the output rate $r_O=s/\gamma_N$ and the input rate $r_I=s/g_I$. However, since

this relation is not linear, we adopt a transformed version first proposed by Melander *et al.* [14], which depicts the relation between the ratio r_I/r_O and r_I. Denoting this rate response curve by $\tilde{\mathcal{F}}(r_I)$, we have

$$\tilde{\mathcal{F}}(r_I) = \frac{r_I}{r_O} = \frac{\gamma_N(g_I, s)}{g_I}. \quad (14)$$

This transformed version of the rate response curve is also piece-wise linear. It is easy to see that the first turning point in the rate curve is $s/\alpha_1 = A_p$ and that the rate curve in the input rate range $(0, s/\alpha_2)$ can be expressed as

$$\tilde{\mathcal{F}}(r_I) = \begin{cases} 1 & r_I \le A_{\mathcal{P}} \\ \dfrac{\lambda_b + r_I}{C_b} & \dfrac{s}{\alpha_2} \ge r_I \ge A_{\mathcal{P}} \end{cases}. \quad (15)$$

Finally, it is also important to notice that the rate response curve $\tilde{\mathcal{F}}(r_I)$ does not depend on the probing packet size s. This is because, for any given input rate r_I, both $\gamma_N(g_I, s)$ and g_I are proportional to s. Consequently, the ratio between these two terms remains a constant for any s.

2.5 Examples

We use a simple example to illustrate the properties of the fluid response curves. Suppose that we have a 3-hop path with equal capacity $C_i = 10$mb/s, $i = 1, 2, 3$. We consider two routing matrices and flow rate settings that lead to the same link load at each hop.

In the first setting, the flow rate vector $\mathbf{x} = (4, 7, 8)$ and the routing pattern is *one-hop* persistent, i.e., $\mathbf{R} = \text{diag}(1, 1, 1)$. In the second setting, the flow rate vector $\mathbf{x} = (4, 3, 1)$ and the routing pattern is *path* persistent. That is,

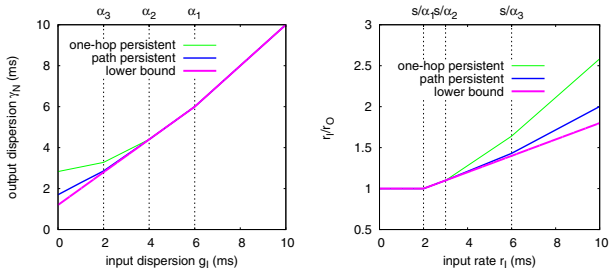
$$\mathbf{R} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \quad (16)$$

Both of the settings result in the same path configuration matrix

$$\mathbf{H} = \begin{pmatrix} 10 & 10 & 10 \\ 4 & 7 & 8 \end{pmatrix}. \quad (17)$$

The probing packet size s is 1500 bytes. The fluid gap response curves for the two routing patterns are plotted in Fig. 1(a). In this example, both curves have 4 linear segments separated by turning points $\alpha_1 = 6$ms, $\alpha_2 = 4$ms, and $\alpha_3 = 2$ms. Note that part of the curve for path-persistent routing appears below the one for one-hop persistent routing. The lower bound $\mathcal{S}$ identified in Property 4 is also plotted in the figure. This lower bound is the gap response curve of the single-hop path comprising only the tight link L_3.

The rate response curves for the two examples are given in Fig. 1(b), where the three turning points are 2mb/s, 3mb/s, and 6mb/s respectively. Due to the transformation



(a) gap response curve (b) rate response curve

Figure 1: An example of multi-hop response curves.

we adopted, the rate curve for one-hop persistent routing still remains as an upper bound for the rate curves associated with the other routing patterns. From Fig. 1(b), we also see that, similar to the gap curves, the two multi-hop rate response curves and their lower bound $\tilde{\mathcal{S}}(r_I)$ (i.e., the transformed rate version of $\mathcal{S}(g_I, s)$) share the same first and second linear segments.

2.6 Discussion

We conclude this section by discussing several major challenges in extending the response curve analysis to a multi-hop path carrying bursty cross-traffic flows. First, notice that with bursty cross-traffic, even when the input dispersion and packet-train parameters remain constant, the output dispersion becomes random, rather than deterministic as in fluid cross-traffic. The gap response curve $\mathcal{Z}$, defined as the functional relation between the statistical mean of the output dispersion and the input dispersion, is much more difficult to penetrate than the fluid curve $\mathcal{F}$. Second, unlike in the fluid case, where both packet-train length n and probing packet size s have no impact on the rate response curve $\tilde{\mathcal{F}}(r_I)$, the response curves in bursty cross-traffic are strongly related to these two packet-train parameters. Finally, a full characterization of a fluid flow only requires one parameter – its arrival rate, while a full characterization of a bursty flow requires several stochastic processes. In what follows, we address these problems and extend our analysis to multi-hop paths with bursty cross-traffic.

3 Basics of Non-Fluid Analysis

In this section, we present a stochastic formulation of the multi-hop bandwidth measurement problem and derive a recursive expression for the output dispersion random variable. This expression is a fundamental result that the asymptotic analysis in Section 4 is based upon.

3.1 Formulating Bursty Flows

We keep most of the notations the same as in the previous section, although some of the terms are extended to have a different meaning, which we explain shortly. Since cross-traffic flows now become bursty flows of data packets, we adopt the definitions of several random processes (Definition 1-6) in [9] to characterize them. However, these definitions need to be refined to be specific to a given router and flow aggregation. In what follows, we only give the definitions of two random processes and skip the others. The notations for all six random processes are given in Table 3.1.

Definition 2 *The cumulative traffic arrival process of flow aggregation* $\mathbf{p}$ *at link* L_i*, denoted as* $\{V_i(\mathbf{p},t), 0 \le t < \infty\}$ *is a random process counting the total amount of data (in bits) received by hop* L_i *from flow aggregation* $\mathbf{p}$ *up to time instance* t*.*

Definition 3 *Hop workload process of* L_i *with respect to flow aggregation* $\mathbf{p}$*, denoted as* $\{W_i(\mathbf{p},t), 0 \le t < \infty\}$ *indicates the sum at time instance* t *of service times of all packets in the queue and the remaining service time of the packet in service, assuming that flow aggregation* $\mathbf{p}$ *is the only traffic passing through link* L_i*.*

We next make several modeling assumptions on cross-traffic flows. First, we assume that all flows have stationary arrivals.

Assumption 1 *For any cross-traffic flow* j *that enters the path from link* L_i*, the cumulative traffic arrival process* $\{V_i(\mathbf{f}_j,t)\}$ *has ergodic stationary increments. That is, for any* $\delta > 0$*, the* δ*-interval traffic intensity process* $\{Y_{i,\delta}(\mathbf{f}_j,t)\}$ *is a mean-square ergodic process with time-invariant distribution and ensemble mean* x_j*.*

We explain this assumption in more details. First, the stationary increment assumption implies that the increment process of $\{V_i(\mathbf{f}_j,t)\}$ for any given time interval δ, namely $\{V_i(\mathbf{f}_j,t+\delta) - V_i(\mathbf{f}_j,t) = \delta Y_{i,\delta}(\mathbf{f}_j,t)\}$, has a time-invariant distribution. This further implies that the δ-interval traffic intensity process $\{Y_{i,\delta}(\mathbf{f}_j,t)\}$ is *identically distributed*, whose marginal distribution at any time instance t can be described by the same random variable $Y_{i,\delta}(\mathbf{f}_j)$. Second, the mean-square ergodicity implies that, as the observation interval δ increases, the random variable $Y_{i,\delta}(\mathbf{f}_j)$ converges to x_j in the mean-square sense. In other words, the variance of $Y_{i,\delta}(\mathbf{f}_j)$ decays to 0 as $\delta \to \infty$, i.e.,

$$\lim_{\delta\to\infty} E\left[\left(Y_{i,\delta}(\mathbf{f}_j) - x_j\right)^2\right] = 0. \tag{18}$$

Our next assumption states the independent relationship between different flows that enter path $\mathcal{P}$ at the same link.

$\{V_i(\mathbf{p},t)\}$	Cumulative arrival process at L_i w.r.t. $\mathbf{p}$
$\{Y_{i,\delta}(\mathbf{p},t)\}$	Cross-traffic intensity process at L_i w.r.t. $\mathbf{p}$
$\{W_i(\mathbf{p},t)\}$	Hop workload process at L_i w.r.t. $\mathbf{p}$
$\{D_{i,\delta}(\mathbf{p},t)\}$	Workload-difference process at L_i w.r.t. $\mathbf{p}$
$\{U_i(\mathbf{p},t)\}$	Hop utilization process at L_i w.r.t. $\mathbf{p}$
$\{B_{i,\delta}(\mathbf{p},t)\}$	Available bandwidth process at L_i w.r.t. $\mathbf{p}$

Table 1: Random process notations

Assumption 2 *For any two flows* j *and* l *that enter the path at link* L_i*, the two processes* $\{V_i(\mathbf{f}_j,t)\}$ *and* $\{V_i(\mathbf{f}_l,t)\}$ *are independent. Specifically, for any two time instances* t_1 *and* t_2*, the two random variables* $V_i(\mathbf{f}_j,t_1)$ *and* $V_i(\mathbf{f}_l,t_2)$ *are independent.*

As a consequence of the two assumptions we made, the ergodic stationary property also holds for any flow aggregations at their entering link.

Corollary 1 *For any flow aggregation* $\mathbf{p}$ *that enters the path at link* L_i*, i.e.,* $\mathbf{p} \odot \mathbf{e}_i = \mathbf{p}$*, the process* $\{V_i(\mathbf{p},t)\}$ *has ergodic stationary increments. Consequently, the traffic intensity random variable* $Y_{i,\delta}(\mathbf{p})$ *converges to* $\mathbf{xp}$ *in the mean-square sense*

$$\lim_{\delta\to\infty} E\left[\left(Y_{i,\delta}(\mathbf{p}) - \mathbf{xp}\right)^2\right] = 0. \tag{19}$$

Due to Szczotka [18], [19], the workload process $\{W_i(\mathbf{p},t)\}$ will "inherit" the ergodic stationarity property from the traffic arrival process $\{V_i(\mathbf{p},t)\}$. This property is further carried over to the δ-interval workload-difference process $\{D_{i,\delta}(\mathbf{p},t)\}$ and the available bandwidth process $\{B_{i,\delta}(\mathbf{p},t)\}$. This distributional stationarity allows us to focus on the corresponding random variables $W_i(\mathbf{p})$, $D_{i,\delta}(\mathbf{p})$, and $B_{i,\delta}(\mathbf{p})$. It is easy to get, from their definitions, that the statistical means of $D_{i,\delta}(\mathbf{p})$ and $B_{i,\delta}(\mathbf{p})$ are 0 and $C_i - \mathbf{xp}$, respectively[6]. Further, the ergodicity property leads to the following result.

Lemma 1 *For any flow aggregation* $\mathbf{p}$ *that enter the path at link* L_i*, the random variable* $B_{i,\delta}(\mathbf{p})$ *converges in the mean-square sense to* $C_i - \mathbf{xp}$ *as* $\delta \to \infty$*, i.e.,*

$$\lim_{\delta\to\infty} E\left[\left(B_{i,\delta}(\mathbf{p}) - (C_i - \mathbf{xp})\right)^2\right] = 0. \tag{20}$$

On the other hand, notice that unlike $\{Y_{i,\delta}(\mathbf{p},t)\}$ and $\{B_{i,\delta}(\mathbf{p},t)\}$, the workload-difference process $\{D_{i,\delta}(\mathbf{p},t)\}$ is not a moving average process by nature. Consequently, the mean-square ergodicity of $\{D_{i,\delta}(\mathbf{p},t)\}$ does not cause the variance of $D_{i,\delta}(\mathbf{p})$ to decay with respect to the increase of δ. Instead, we have the following lemma.

Lemma 2 *The variance of the random variable* $D_{i,\delta}(\mathbf{p})$ *converges to* $2Var[W_i(\mathbf{p})]$ *as* δ *increases:*

$$\lim_{\delta\to\infty} E\left[\left(D_{i,\delta}(\mathbf{p}) - 0\right)^2\right] = 2Var\left[W_i(\mathbf{p})\right]. \tag{21}$$

To obtain our later results, not only do we need to know the asymptotic variance of $Y_{i,\delta}(\mathbf{p})$, $D_{i,\delta}(\mathbf{p})$ and $B_{i,\delta}(\mathbf{p})$ when δ approaches infinity, but also we often rely on their variance being uniformly bounded (for any δ) by some constant. This condition can be easily justified from a practical standpoint. First note that cross-traffic arrival rate is bounded by the capacities of incoming links at a given router. Suppose that the sum of all incoming link capacities at hop L_i is C_+, then $Y_{i,\delta}(\mathbf{p})$ is distributed in a finite interval $[0, C_+]$ and its variance is uniformly bounded by the constant C_+^2 for any observation interval δ. Similarly, the variance of $B_{i,\delta}(\mathbf{p})$ is uniformly bounded by the constant C_i^2. The variance of $D_{i,\delta}(\mathbf{p})$ is uniformly bounded by the constant $4Var[W_i(\mathbf{p})]$ for any δ, which directly follows from the definition of $D_{i,\delta}(\mathbf{p})$.

Finally, we remind that some of the notations introduced in Section 2.1 now are used with a different meaning. The rate of the bursty cross-traffic flow j, denoted by x_j, is the probabilistic mean of the traffic intensity random variable $Y_{i,\delta}(\mathbf{f}_j)$, which is also the *long-term average* arrival rate of flow j at any link it traverses. The term $\lambda_i = \mathbf{x}\mathbf{r}_i$ becomes the long-term average arrival rate of the aggregated cross-traffic at link L_i. The term $A_i = C_i - \lambda_i$ is the long-term average hop available bandwidth at link L_i. Again recall that we explicitly target the measurement of long-term averages of available bandwidth and/or cross-traffic intensity, instead of the corresponding metrics in a certain time interval.

3.2 Formulating Packet Train Probing

We now consider an infinite series of packet-trains with input inter-packet dispersion g_I, packet size s, and packet-train length n. This series is driven to path $\mathcal{P}$ by a point process $\Lambda(t) = \max\{m \geq 0 : T_m \leq t\}$ with sufficient large inter-probing separation. Let $d_1(m,i)$ and $d_n(m,i)$ be the departure time instances from link L_i of the first and last probing packets in the m^{th} packet-train. We define the *sampling interval* of the packet-train as the total spacing $\Delta = d_n(m,i) - d_1(m,i)$, and the *output dispersion* as the average spacing $G = \Delta/(n-1)$ of the packet-train. Both Δ and G are random variables, whose statistics might depend on several factors such as the input dispersion g_I, the packet-train parameters s and n, the packet-train index m in the probing series, and the hop L_i that the output dispersion G is associated with. Therefore, a full version of G is written as $G_i(g_I, s, n, m)$. However, for notation brevity, we often omit the parameters that have little relevance to the topic under discussion.

We now formally state the questions we address in this paper. Note that a realization of the stochastic process $\{G_N(g_I, s, n, m), 1 \leq m < \infty\}$ is just a packet-train probing experiment. We examine the sample-path time-average of this process and its relationship to g_I when keeping s and n constant. This relationship, previously denoted by $\mathcal{Z}$, is called the gap response curve of path $\mathcal{P}$.

Notice that the ergodic stationarity of cross-traffic arrival, as we assumed previously, can reduce our response curve analysis to the investigation of a single random variable. This is because each packet-train comes to see a multi-hop system of the same stochastic nature and the output dispersion process $\{G_N(m), 1 \leq m < \infty\}$ is an *identically distributed* random sequence, which can be described by the output dispersion random variable G_N. The sample-path time average of the output dispersion process coincides with the mean of the random variable G_N[7]. Therefore, in the rest of the paper, we focus on the statistics of G_N and drop the index m.

In our later analysis, we compare the gap response curve of $\mathcal{P}$ with that of the *fluid counterpart* of $\mathcal{P}$ and prove that the former is lower-bounded by the latter.

Definition 4 *Suppose that path $\mathcal{P}$ has a routing matrix $\mathbf{R}$ and a flow rate vector $\mathbf{x}$ and that path $\tilde{\mathcal{P}}$ has a routing matrix $\tilde{\mathbf{R}}$ and a flow rate vector $\tilde{\mathbf{x}}$. $\tilde{\mathcal{P}}$ is called the fluid counterpart of $\mathcal{P}$ if 1) all cross-traffic flows traversing $\tilde{\mathcal{P}}$ are constant-rate fluid; 2) the two paths $\tilde{\mathcal{P}}$ and $\mathcal{P}$ have the same configuration matrix; and 3) there exists a row-exchange matrix T, such that $T\mathbf{R} = \tilde{\mathbf{R}}$ and $T\mathbf{x} = \tilde{\mathbf{x}}$.*

From this definition, we see that for every flow j in $\mathcal{P}$, there is a corresponding fluid flow j' in the fluid counterpart of $\mathcal{P}$ such that flow j' have the same average intensity and routing pattern as those of flow j. Note that the third condition in Definition 4 is made to allow the two flows have different indices, i.e., to allow $j \neq j'$.

A second focus of this paper is to study the impact of packet-train parameters s and n on the response curves. That is, for any given input rate r_I and other parameters fixed, we examine the convergence properties of the output dispersion random variable $G_N(s/r_I, s, n)$ as s or n tends to infinity.

3.3 Recursive Expression of G_N

We keep input packet-train parameters g_I, s, and n constant and next obtain a basic expression for the output dispersion random variable G_N.

Lemma 3 *Letting $G_0 = g_I$, the random variable G_i has the following recursive expression*

$$\begin{aligned} G_i &= \sum_{k=1}^{i} \frac{Y_{k,\Delta_{k-1}}(\Gamma_{k,i}) G_{k-1}}{C_i} + \frac{s}{C_i} + \frac{\tilde{I}_i}{n-1} \\ &= G_{i-1} + \frac{D_{i,\Delta_{i-1}}(\mathbf{e}_i)}{n-1} + \frac{R_i}{n-1}, \end{aligned} \tag{22}$$

where the term R_i is a random variable representing the extra queuing delay[8] (besides the queuing delay caused by

the workload process $\{W_i(\mathbf{e}_i, t)\}$) experienced at L_i by the last probing packet in the train. The term $\tilde{I}_i$ is another random variable indicating the hop idle time of L_i during the sampling interval of the packet train.

This result is very similar to Lemma 5 in [9]. However, due to the random input packet-train structure at L_i, all but the term s/C_i in (22) become random variables. Some terms, such as $D_{i,\Delta_{i-1}}(\mathbf{e}_i)$ and $Y_{k,\Delta_{k-1}}(\Gamma_{k,i})$, even have two dimensions of randomness. To understand the behavior of probing response curves, we need to investigate the statistical properties of each term in (22).

4 Response Curves in Bursty Cross-Traffic

In this section, we first show that the gap response curve $\mathcal{Z} = E[G_N(g_I, s, n)]$ of a multi-hop path $\mathcal{P}$ is lower bounded by its fluid counterpart $\mathcal{F} = \gamma_N(g_I, s)$. We then investigate the impact of packet-train parameters on $\mathcal{Z}$.

4.1 Relation Between $\mathcal{Z}$ and $\mathcal{F}$

Our next lemma shows that passing through a link can only increase the dispersion random variable in mean.

Lemma 4 *For $1 \le i \le N$, the output dispersion random variable G_i has a mean no less than that of G_{i-1}. That is, $E[G_i] \ge E[G_{i-1}]$.*

Using the first part of (22), our next lemma shows that for any link L_i, the output dispersion random variable G_i is lower bounded in mean by a linear combination of the output dispersion random variables G_k, where $k < i$.

Lemma 5 *For $1 \le i \le N$, the output dispersion random variable G_i satisfies the following inequality*

$$E[G_i] \ge \frac{1}{C_i}\left(\sum_{k=1}^{i} \mathbf{x}\Gamma_{k,i} E[G_{k-1}] + s\right). \tag{23}$$

From Lemma 4 and Lemma 5, we get

$$E[G_i] \ge \max\left(E[G_{i-1}], \frac{\sum_{k=1}^{i} \mathbf{x}\Gamma_{k,i} E[G_{k-1}] + s}{C_i}\right). \tag{24}$$

This leads to the following theorem.

Theorem 2 *For any input dispersion g_I, packet-train parameters s and n, the output dispersion random variable G_N of path $\mathcal{P}$ is lower bounded in mean by the output dispersion $\gamma_N(g_I, s)$ of the fluid counterpart of $\mathcal{P}$:*

$$E[G_N(g_I, s, n)] \ge \gamma_N(g_I, s). \tag{25}$$

Proof: We apply mathematical induction to i. When $i = 0$, $E[G_0] = \gamma_0 = g_I$. Assuming that (25) holds for $0 \le i < N$, we next prove that it also holds for $i = N$. Recalling (24), we have

$$\begin{aligned} E[G_N] &\ge \max\Big(E[G_{N-1}], \frac{\sum_{k=1}^{N} \mathbf{x}\Gamma_{k,N} E[G_{k-1}] + s}{C_N}\Big) \\ &\ge \max\Big(\gamma_{N-1}, \frac{\sum_{k=1}^{N} \mathbf{x}\Gamma_{k,N} \gamma_{k-1} + s}{C_N}\Big) = \gamma_N, \end{aligned}$$

where the second inequality is due to the induction hypothesis, and the last equality is because of Theorem 1. ■

Theorem 2 shows that in the entire input gap range, the piece-wise linear fluid gap response curve $\mathcal{F}$ discussed in Section 2 is a lower bound of the real gap curve $\mathcal{Z}$. The deviation between the real curve $\mathcal{Z}$ and its fluid lower bound $\mathcal{F}$, which is denoted by $\beta_N(g_I, s, n)$ or β_N for short, can be recursively expressed in the following, where we let $\beta_0 = 0$:

$$\beta_i = \begin{cases} \beta_{i-1} + \dfrac{E[R_i]}{n-1} & \gamma_i = \gamma_{i-1} \\ \dfrac{1}{C_i}\sum_{k=1}^{i} \mathbf{x}\Gamma_{k,i}\beta_{k-1} + \dfrac{E[\tilde{I}_i]}{n-1} & \gamma_i > \gamma_{i-1} \end{cases}. \tag{26}$$

In what follows, we study the asymptotics of the curve deviation β_N when input packet-train parameters s or n becomes large and show that the fluid lower bound $\mathcal{F}$ is in fact a *tight* bound of the real response curve $\mathcal{Z}$.

4.2 Impact of Packet Train Parameters

We now demonstrate that for any input probing rate r_I, the curve deviation $\beta_N(s/r_I, s, n)$ vanishes as probing packet size s approaches infinity. We prove this result under the condition of one-hop persistent cross-traffic routing. We also justify this conclusion informally for arbitrary cross-traffic routing and point out the major difficulty in obtaining a rigorous proof. First, we make an additional assumption as follows.

Assumption 3 *Denoting by $P_{i,\delta}(x)$ the distribution function of the δ-interval available bandwidth process $\{B_{i,\delta}(\mathbf{e}_i, t)\}$, we assume that for all $1 \le i \le N$, the following holds*

$$\begin{cases} P_{i,\delta}(r) = o\left(\dfrac{1}{\delta^2}\right) & r < C_i - \mathbf{x}\mathbf{e}_i \\ P_{i,\delta}(r) = 1 - o\left(\dfrac{1}{\delta^2}\right) & r > C_i - \mathbf{x}\mathbf{e}_i \end{cases}. \tag{27}$$

Recall that the mean-square ergodicity assumption we made earlier implies that as the observation interval δ gets large, the random variable $B_{i,\delta}(\mathbf{e}_i)$ converges in distribution to $C_i - \mathbf{x}\mathbf{e}_i$. Assumption 3 further ensures that this convergence is *fast* in the sense of (27). Even though this

condition appears cryptic at first, it is valid in a broad range of cross-traffic environments. The next theorem shows the validity of this assumption under the condition of regenerative[9] link utilization.

Theorem 3 *When hop utilization process $\{U_i(\mathbf{e}_i, t)\}$ is regenerative, condition (27) holds.*

Note that regenerative queue is very common both in practice and in stochastic modeling literature. In fact, all the four traffic types used in [9] lead to regenerative hop workload and consequently lead to regenerative link utilization. We also conjecture that (27) holds under a much milder condition, but we leave its identification as future work.

Our next theorem states formally the convergence property of the output dispersion random variable $G_N(s/r_I, s, n)$ when s increases.

Theorem 4 *Given one-hop persistent cross-traffic routing and the three assumptions made in the paper, for any input rate r_I, the output dispersion random variable G_N of path $\mathcal{P}$ converges in mean to its fluid lower bound γ_N:*

$$\lim_{s\to\infty} E\left[G_N\left(\frac{s}{r_I}, s, n\right) - \gamma_N\left(\frac{s}{r_I}, s\right)\right] = 0. \quad (28)$$

The asymptotic variance of G_N when s increases is upper bounded by some constant K_N:

$$\lim_{s\to\infty} E\left[\left(G_N\left(\frac{s}{r_I}, s, n\right) - \gamma_N\left(\frac{s}{r_I}, s\right)\right)^2\right] \le K_N. \quad (29)$$

Note that the bounded variance, as stated in (29), is an inseparable part of the whole theorem. This is because Theorem 4 is proved using mathematical induction, where the mean convergence of G_N to γ_N can be obtained only when the mean of G_{N-1} converges to γ_{N-1} and when the variance of G_{N-1} remains bounded, as probing packet size $s \to \infty$.

We further point out that by assuming one-hop persistent cross-traffic routing, we have avoided analyzing the departure processes of cross-traffic flows. When a traversing flow of link L_i enters the path from some upstream link of L_i, the arrival process of the flow at L_i is its departure process at L_{i-1}. Unfortunately, in the queueing theory literature, there is no exact result for departure processes in FCFS queueing models if one goes beyond the assumption of Poisson arrivals. Motivated by the intractability of this problem, researchers have focused their attentions on approximations [12], [15].

To accommodate arbitrary cross-traffic routing patterns, we also need an approximation assumption which says that any cross-traffic flow that traverses link L_i (regardless of wether it enters the path from L_i or some upstream link of L_i) exhibits ergodic stationary arrival at L_i. Under this assumption, which we call "stationary departure approximation," it becomes easy to extend Theorem 4 to cover arbitrary cross-traffic routing patterns. We skip the details of this step and next apply the stationary departure approximation to examine the impact of packet-train length n on the response curve $\mathcal{Z}$.

Theorem 5 *Under the first two assumptions and the "stationary departure approximation", for any N-hop path $\mathcal{P}$ with arbitrary cross-traffic routing, for any input dispersion $g_I \in (0, \infty)$ and any probing packet size s, the random variable G_N converges to its fluid lower bound γ_N in the mean-square sense as $n \to \infty$,*

$$\lim_{n\to\infty} E\left[\left(G_N(g_I, s, n) - \gamma_N(g_I, s)\right)^2\right] = 0. \quad (30)$$

Let us make several comments on the conditions of this result. First note that Assumption 3 is not necessary in this theorem. Also notice that in a single-hop path (i.e., $N = 1$), the theorem can be proved without the stationary departure approximation. However, in the multi-hop cases, the approximation is needed even when cross-traffic routing is one-hop persistent. The reason is that when n is large, the probing packet-train is also viewed as a flow, whose arrival characteristics at all but the first hop are addressed by the stationary departure approximation.

Theorem 5 shows that when the packet-train length n increases while keeping s constant, not only $E[G_N]$ converges to its fluid bound γ_N, but also the variance of G_N decays to 0. This means that we can expect almost the same output dispersion in different probings.

4.3 Discussion

Among the assumptions in this paper, some are critical in leading to our results while others are only meant to simplify discussion. We point out that the distributional stationarity assumption on cross-traffic arrivals can be greatly relaxed without harming our major results. However, this comes at the expense of much more intricate derivations. This is because when cross-traffic arrivals are allowed to be only second-order stationary or even non-stationary, the output dispersion process $\{G_N(m)\}$ will no longer be identically distributed. Consequently, the analysis of probing response curves cannot be reduced to the investigation of a *single* output dispersion random variable. Moreover, we also have to rely on an ASTA assumption on packet-train probing [9] to derive the results in this paper, which we have avoided in the present setting.

Also note that the inter-flow independence assumption is made to maintain the distributional stationarity of cross-traffic arrivals at a flow aggregation level. It only helps us

avoid unnecessary mathematical rigor and is insignificant in supporting our major conclusions.

On the other hand, the mean-square ergodicity plays a central role in the (omitted) proofs for Theorem 4 and Theorem 5. A cross-traffic flow with mean-square ergodicity, when observed in a large timescale, has an almost constant arrival rate. This "asymptotically fluid like" property, is very common among the vast majority of traffic models in stochastic literature, and can be decoupled from any type of traffic stationarity. Consequently, our results have a broad applicability in practice.

Next, we provide experimental evidence for our theoretical results using testbed experiments and real Internet measurement data.

5 Experimental Verification

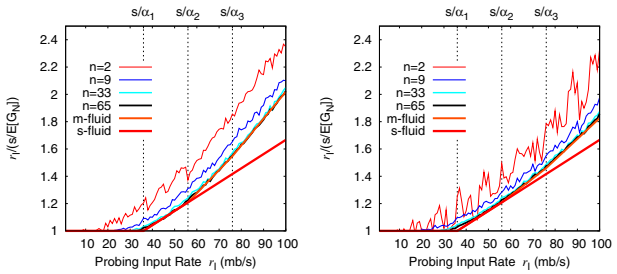
In this section, we measure the response curves in both testbed and real Internet environments. The results not only provide experimental evidence to our theory, but also give quantitative ideas of the curve deviation given in (26). To obtain the statistical mean of the probing output dispersions, we rely on direct measurements using a number of probing samples. Even though this approach can hardly produce a smooth response curve, the bright side is that it allows us to observe the output dispersion variance, reflected by the degree of smoothness of the measured response curve.

5.1 Testbed Experiments

In our first experiment, we measure in the Emulab testbed [1] the response curves of a three-hop path with the following configuration matrix (all in mb/s) and one-hop persistent cross-traffic routing

$$\mathbf{H} = \begin{pmatrix} 96 & 96 & 96 \\ 20 & 40 & 60 \end{pmatrix}. \tag{31}$$

We generate cross-traffic using three NLANR [2] traces. All inter-packet delays in each trace are scaled by a common factor so that the average rate during the trace duration becomes the desired value. The trace durations after scaling are 1-2 minutes. We measure the average output dispersions at 100 input rates, from 1mb/s to 100mb/s with 1mb/s increasing step. For each input rate, we use 500 packet-trains with packet size 1500 bytes. The packet train length n is 65. The inter-probing delay is controlled by a random variable with sufficiently large mean. The whole experiment lasts for about 73 minutes. All three traffic traces are replayed at random starting points once the previous round is finished. By recycling the same traces in this fashion, we make the cross-traffic last until the experiment ends without creating periodicity. Also note that the packet-trains are injected with their input rates so arranged that the 500 trains



(a) one-hop persistent routing (b) path-persistent routing

Figure 2: Measured response curves using different packet train-length in the Emulab testbed.

for each input rate is evenly separated during the whole testing period.

This experiment not only allows us to measure the response curve for $n = 65$, but also for any packet-train length k such that $2 \le k < n = 65$, by simply taking the dispersions of the first k packets in each train. Fig. 2(a) shows the rate response curve $\tilde{\mathcal{Z}}(r_I, s, n)$ for $k = 2, 9, 33$ and 65 respectively. For comparison purposes, we also plot in the figure the multi-hop fluid curve $\tilde{\mathcal{F}}(r_I)$, computed from Theorem 1, and the single-hop fluid curve $\tilde{\mathcal{S}}(r_I)$ of the tight link L_3. The rate response curves $\tilde{\mathcal{Z}}(r_I, s, n)$ is defined as follows

$$\tilde{\mathcal{Z}}(r_I, s, n) = \frac{r_I}{s/E[G_N(s/r_I, s, n)]}. \tag{32}$$

First note that the multi-hop fluid rate curve comprises four linear segments separated by turning points 36mb/s, 56mb/s, and 76mb/s. The last two linear segments have very close slopes and they are not easily distinguishable from each other in the figure. We also clearly see that the rate curve asymptotically approaches its fluid lower bound as packet-train length n increases. The curves for $n = 33$ and $n = 65$ almost coincide with the fluid bound. Also note that the smoothness of the measurement curve reflects the variance of the output dispersion random variables. As the packet train length increases, the measured curve becomes smoother, indicating the fact that the variance of the output dispersions is decaying. These observations are all in agreement with those stated in Theorem 5.

Unlike single-hop response curves, which have no deviation from the fluid bound when the input rate r_I is greater than the link capacity, multi-hop response curves usually deviate from its fluid counterpart in the entire input range. As we see from Fig. 2(a), even when the input rate is larger than 96mb/s, the measured curves still appear above $\tilde{\mathcal{F}}$. Also observe that the single-hop fluid curve $\tilde{\mathcal{S}}$ of the tight link L_3 coincides with the multi-hop fluid curve $\tilde{\mathcal{F}}$ within the input rate range $(0, 56)$ but falls below $\tilde{\mathcal{F}}$ in the input rate range $(56, \infty)$.

Finally, we explain why we choose the link capacities to

be 96mb/s instead of the fast ethernet capacity 100mb/s. In fact, we did set the link capacity to be 100mb/s. However, we noticed that the measured curves can not get arbitrarily close to their fluid bound $\tilde{\mathcal{F}}$ computed based on the fast ethernet capacity. Using pathload to examine the true capacity of each Emulab link, we found that their IP layer capacities are in fact 96mb/s, not the same as their nominal value 100mb/s.

In our second experiment, we change the cross-traffic routing to path-persistent while keeping the path configuration matrix the same as given by (31). Therefore, the flow rate vector now becomes $(20, 20, 20)$.

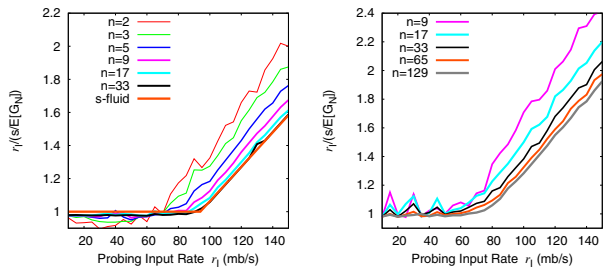
We repeat the same packet-train probing experiment and the results are plotted in Fig. 2(b). The multi-hop fluid rate curve $\tilde{\mathcal{F}}$ still coincides with $\tilde{\mathcal{S}}$ in the input rate range $(0, 56)$. When input rate is larger than 56mb/s, the curve $\tilde{\mathcal{F}}$ positively deviates from $\tilde{\mathcal{S}}$. However, the amount of deviation is smaller than that in one-hop persistent routing. The measured curve approaches the fluid lower bound $\tilde{\mathcal{F}}$ with decaying variance as packet-train length increases. For $n = 33$ and $n = 65$, the measured curves become hardly distinguishable from $\tilde{\mathcal{F}}$.

We have conducted experiments using paths with more hops, with more complicated cross-traffic routing patterns, and with various path configurations. Furthermore, we examined the impact of probing packet size using ns2 simulations, where the packet size can be set to any large values. Results obtained (not shown for brevity) all support our theory very well.

5.2 Real Internet Measurements

We conducted packet-train probing experiments on several Internet paths in the RON testbed to verify our analysis in real networks. Since neither the path configuration nor the cross-traffic routing information is available for these Internet paths, we are unable to provide the fluid bounds. Therefore, we verify our theory by observing the convergence of the measured curves to a piece-wise linear curve as packet-train length increases.

In the first experiment, we measure the rate response curve of the path from the RON node lulea in Sweden to the RON node at CMU. The path has 19 hops and a fast-ethernet minimum capacity, as we find out using traceroute and pathrate. We probe the path at 29 different input rates, from 10mb/s to 150mb/s with a 5mb/s increasing step. For each input rate, we use 200 packet-trains of 33 packets each to estimate the output probing rate $s/E[G_N]$. The whole experiment takes about 24 minutes. Again, the 200 packet-trains for each of the 29 input rates are so arranged that they are approximately evenly separated during the 24-minute testing period. The measured rate response curves associated with packet-train length 2, 3, 5, 9, 17, and 33 are plotted in Fig. 3(a), where we see that the response



(a) lulea → CMU (b) pwh → NYU

Figure 3: Measured response curves of two Internet paths in RON testbed .

curve approaches a piece-wise linear bound as packet-train length increases. At the same time, response curves measured using long trains are smoother than those measured using short trains, indicating the decaying variance of output dispersions. In this experiment, the curve measured using probing trains of 33-packet length exhibits sufficient smoothness and clear piece-wise linearity. We have observed two linear segments from the figure. A further investigation shows that the fluid bound of this 19-hop path only has two linear segments.

Based on (15), we apply linear regression on the second linear segment to compute the capacity C_b and the cross-traffic intensity λ_b of the tight link and get $C_b = 96$mb/s and $\lambda_b = 2$mb/s. Using these results, we retroactively plot the single-hop fluid bounds and observe that it almost overlaps with the measured curve using packet-trains of 33-packet length. Notice that the bottleneck link is under very light utilization during our 24-minute measurement period. We can also infer based on our measurement that the available bandwidth of the path is constrained mainly by the capacity of the bottleneck link and that the probing packet-trains have undergone significant interaction with cross-traffic at non-bottleneck links. Otherwise, according to Theorem 3 in [9], the response curves measured using short train lengths would not have appeared above the single-hop fluid bound when the input rate is larger than the tight link capacity 96mb/s. We believe that the tight link of the path is one of the last-mile lightly utilized fast-ethernet links and that the backbone links are transmitting significant amount of cross-traffic even though they still have available bandwidth much more than the fast-ethernet capacity. Also notice that similar to our testbed experiments, fast-ethernet links only have 96mb/s IP-layer capacity.

We repeat the same experiment on another path from the RON node pwh in Sunnyvale California to the NYU RON node. This path has 13 hops and a fast-ethernet minimum capacity. Due to substantial cross-traffic burstiness along the path, we use packet-trains of 129-packet length in our probing experiment. The other parameters such as the input rates and the number of trains used for each rate are

the same as in the previous experiment. The whole measurement duration is about 20 minutes. The measured response curves are plotted in Fig. 3(b). As we see, the results exhibit more measurement variability compared to the lulea→CMU path. However, as packet-train length increases, the variability is gradually smoothed out and the response curve converges to a piece-wise linear bound. We again apply linear regression on the response curve with packet-train length 129 to obtain the tight link information. We get $C_b = 80$mb/s and $\lambda_b = 3$mb/s, which does not agree with the minimum capacity reported by pathrate. We believe that pathrate reported the correct information. Our underestimation is most probably due to the fact that there are links along the path with very similar available bandwidth. Consequently, the second linear segment become too short to detect. The linear segment we are acting upon is likely to be a latter one. This experiment confirms our analysis, at the same time shows some of the potential difficulties in exacting tight link information from the response curves.

6 Implications

We now discuss the implications of our results on existing measurement proposals. Except for pathChirp, all other techniques such as TOPP, pathload, PTR, and Spruce are related to our analysis.

6.1 TOPP

TOPP is based on multi-hop fluid rate response curve $\tilde{\mathcal{F}}$ with one-hop persistent cross-traffic routing. TOPP uses packet-pairs to measure the real rate response curve $\tilde{\mathcal{Z}}$, and assumes that the measured curve will be the same as $\tilde{\mathcal{F}}$ when a large number of packet-pairs are used. However, our analysis shows that the real curve $\tilde{\mathcal{Z}}$ is different from $\tilde{\mathcal{F}}$, especially when packet-trains of short length are used (e.g., packet-pairs). Note that there is not much path information in $\tilde{\mathcal{Z}}$ that is readily extractable unless it is sufficiently close to its fluid counterpart $\tilde{\mathcal{F}}$. Hence, to put TOPP to work in practice, one must use long packet-trains instead of packet-pairs.

6.2 Spruce

Using the notations in this paper, we can write spruce's available bandwidth estimator as follows

$$C_b\left(1 - \frac{G_N(s/C_b, s, n) - s/C_b}{s/C_b}\right), \qquad (33)$$

where the probing packet size s is set to 1500bytes, the packet-train length $n = 2$, and the bottleneck link capacity C_b is assumed known.

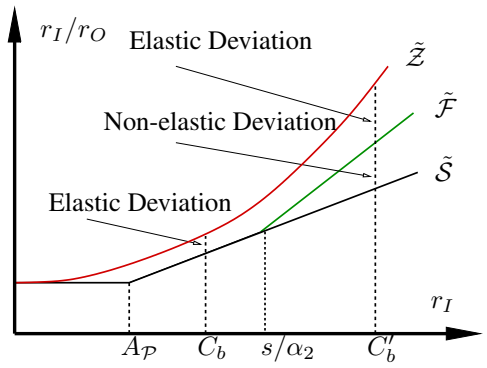


Figure 4: Illustration of two types of curve deviations.

It is shown in [9] that the spruce estimator is unbiased in single-hop paths regardless of the packet-train parameters s and n. This means that the statistical mean of (33) is equal to $A_{\mathcal{P}}$ for any $s > 0$ and any $n \geq 2$. In a multi-hop path $\mathcal{P}$, a necessary condition to maintain the unbiasedness property of the spruce estimator is

$$\tilde{\mathcal{Z}}(C_b, s, n) = \frac{\lambda_b + C_b}{C_b} = \tilde{\mathcal{S}}(C_b). \qquad (34)$$

This means that at the input rate point C_b, the real rate response of path $\mathcal{P}$ must be equal to the single-hop fluid rate response at the tight link of $\mathcal{P}$.

This condition is usually not satisfied. Instead, due to Theorem 2 and Property 4, we have

$$\tilde{\mathcal{Z}}(C_b, s, n) \geq \tilde{\mathcal{F}}(C_b) \geq \tilde{\mathcal{S}}(C_b). \qquad (35)$$

This implies that (33) is a negatively biased estimator of $A_{\mathcal{P}}$. The amount of bias is given by

$$C_b\Big(\tilde{\mathcal{Z}}(C_b, s, n) - \tilde{\mathcal{F}}(C_b)\Big) + C_b\Big(\tilde{\mathcal{F}}(C_b) - \tilde{\mathcal{S}}(C_b)\Big). \qquad (36)$$

The first additive term in (36) is the measurement bias caused by the curve deviation of $\tilde{\mathcal{Z}}$ from $\tilde{\mathcal{F}}$ at input rate C_b, which vanishes as $n \to \infty$ due to Theorem 5. Hence we call it *elastic bias*. The second additive term is the portion of measurement bias caused by the curve deviation of $\tilde{\mathcal{F}}$ from $\tilde{\mathcal{S}}$ at input rate C_b, which remains constant with respect to the packet-train parameters s and n. Therefore it is *non-elastic*. We illustrate the two types of curve deviations in Fig. 4. Note that when $C_b < s/\alpha_2$, non-elastic bias is 0. Further recall that $s/\alpha_2 \geq A_{b2}$ as stated in Property 3. Hence, a sufficient condition for zero non-elastic bias is $C_b \leq A_{b2}$. Conceptually, elastic deviation stems from cross-traffic burstiness and non-elastic deviation is a consequence of multi-hop effects.

In Table 2, we give the amount measurement bias caused by the two types of curve deviations in both the Emulab testbed experiments and the real Internet probing measurement on the path from lulea to CMU. Note that in the

experiment	elastic bias	non-elastic bias	total bias
Emulab-1	0.56×96	0.315×96	74.4
Emulab-2	0.28×96	0.125×96	38.8
lulea-cmu	0.25×96	0	24

Table 2: Spruce bias in Emulab and Internet experiment (in mb/s).

testbed experiment using a 3-hop path with one-hop persistent routing, spruce suffers about 74mb/s measurement bias, which is twice as much as the actual path available bandwidth 36mb/s. In the second Emulab experiment using path-persistent cross-traffic, the measurement bias is reduced to 38.8mb/s, which however is still more than the actual available bandwidth. In both cases, spruce estimator converges to negative values. We used spruce to estimate the two paths and it did in fact give 0mb/s results in both cases. For the Internet path from lulea to CMU, spruce suffers 24mb/s negative bias and produces a measurement result less than 70mb/s, while the real value is around 94mb/s. We also use pathload to measure the three paths and observe that it produces pretty accurate results.

The way to reduce elastic-bias is to use long packet-trains instead of packet-pairs. In the lulea→CMU experiment, using packet-trains of 33-packet, spruce can almost completely overcome the 24mb/s bias and produce an accurate result. However, there are two problems of using long packet-trains. First, there is not a deterministic train length that guarantees negligible measurement bias on any network path. Second, when router buffer space is limited and packet-train length are too large, the later probing packets in each train may experience frequent loss, making it impossible to accurately measure $\tilde{\mathcal{F}}(C_b)$. After all, spruce uses input rate C_b, which can be too high for the bottleneck router to accommodate long packet-trains. On the other hand, note that non-elastic bias is an inherit problem for spruce. There is no way to overcome it by adjusting packet-train parameters.

6.3 PTR and pathload

PTR searches the first turning point in the response curve $\tilde{\mathcal{Z}}(r_I, s, n)$ and takes the input rate at the turning point as the path available bandwidth $A_{\mathcal{P}}$. This method can produce accurate result when the real response curve $\tilde{\mathcal{Z}}$ is close to $\tilde{\mathcal{F}}$, which requires packet-train length n to be sufficiently large. Otherwise, PTR is also negatively biased and underestimates $A_{\mathcal{P}}$. The minimum packet-train length needed is dependent on the path conditions. The current version of PTR use packet train length $n = 60$, which is probably insufficient for the Internet path from pwh to CMU experimented in this paper.

Pathload is in spirit similar to PTR. However, it searches the available bandwidth region by detecting one-way-delay increasing trend within a packet-train, which is different from examining whether the rate response $\tilde{\mathcal{Z}}(r_I, s, n)$ is greater than one [7]. However, since there is a strong statistical correlation between a high rate response $\tilde{\mathcal{Z}}(r_I, s, n)$ and the one-way-delay increasing tend within packet-trains, our analysis can explain the behavior of pathload to a certain extent. Recall that, as reported in [6], pathload underestimates available bandwidth when there are multiple tight links along the path. Our results demonstrate that the deviation of $\tilde{\mathcal{Z}}(r_I, s, n)$ from $\tilde{\mathcal{F}}$ in the input rate range $(0, A_{\mathcal{P}})$ gives rise to a potential underestimation in pathload. The underestimation is maximized and becomes clearly noticeable when non-bottleneck links have the same available bandwidth as $A_{\mathcal{P}}$, given that the other factors are kept the same.

Even through multiple tight links cause one-way-delay increasing trend for packet-trains with input rate less than $A_{\mathcal{P}}$, this is *not* an indication that the network can not sustain such an input rate. Rather, the increasing trend is a *transient* phenomenon resulting from probing intrusion residual, and it disappears when the input packet-train is sufficiently long. Hence, it is our new observation that by further increasing the packet-train length, the underestimation in pathload can be mitigated.

7 Related Work

Besides the measurement techniques we discussed earlier, Melander *et al.* [13] first discussed the rate response curve of a multi-hop network path carrying fluid cross-traffic with one-hop persistent routing pattern. Dovrolis *et al.* [3], [4] considered the impact of cross-traffic routing on the output dispersion rate of a packet-train. It was also pointed out that the output rate of a back-to-back input packet-train (input rate $r_I = C_1$, the capacity of the first hop L_1) converges to a point they call "asymptotic dispersion rate (ADR)" as packet-train length increases. The authors provided an informal justification as to why ADR can be computed using fluid cross-traffic. They demonstrated the computation of ADR for several special path conditions. Note that using the notations in this paper, ADR can be expressed as

$$\lim_{n\to\infty} \frac{s}{G_N(s/C_1, s, n)} = \frac{s}{\gamma_N(s/C_1, s)}. \qquad (37)$$

Our work not only formally explains previous findings, but also generalizes them to such an extent that allows any input rate and any path conditions.

Kang *et al.* [8] analyzed the gap response of a single-hop path with bursty cross-traffic using packet-pairs. The paper had a focus on large input probing rate. Liu *et al.* extended the single-hop analysis for packet-pairs [11] and packet-trains [9] to arbitrary input rates and discussed the impact of packet-train parameters.

8 Conclusion

This paper provides a stochastic characterization of packet-train bandwidth estimation in a multi-hop path with arbitrarily routed cross-traffic flows. Our main contributions include derivation of the multi-hop fluid response curve as well as the real response curve and investigation of the convergence properties of the real response curve with respect to packet-train parameters. The insights provided in this paper not only help understand and improve existing techniques, but may also lead to a new technique that measures tight link capacity.

There are a few unaddressed issues in our theoretical framework. In our future work, we will identify how various factors, such as path configuration and cross-traffic routing, affect the amount of deviation between $\mathcal{Z}$ and $\mathcal{F}$. We are also interested in investigating new approaches that help detect and eliminate the measurement bias caused by bursty cross-traffic in multi-hop paths.

Acknowledgements

Dmitri Loguinov was supported by NSF grants CCR-0306246, ANI-0312461, CNS-0434940.

References

[1] Emulab. http://www.emulab.net.

[2] National Laboratory for Applied Network Research. http://www.nlanr.net.

[3] C. Dovrolis, P. Ramanathan, and D. Moore, "What Do Packet Dispersion Techniques Measure?," *IEEE INFOCOM*, April 2001.

[4] C. Dovrolis, P. Ramanathan, and D. Moore, "Packet Dispersion Techniques and a Capacity Estimation Methodology," *IEEE/ACM Transaction on Networking*, March 2004.

[5] N. Hu and P. Steenkiste, "Evaluation and Characterization of Available Bandwidth Probing Techniques," *IEEE JSAC Special Issue in Internet and WWW Measurement, Mapping, and Modeling*, 3rd Quarter 2003.

[6] M. Jain and C. Dovrolis, "End-to-end available bandwidth: measurement methodology, dynamics, and relation with TCP throughput," *ACM SIGCOMM*, August 2002.

[7] M. Jain and C. Dovrolis, "Ten Fallacies and Pitfalls in End-to-End Available Bandwidth Estimation," *ACM IMC*, October 2004.

[8] S. Kang, X. Liu, M. Dai, and D. Loguinov, "Packet-pair Bandwidth Estimation: Stochastic Analysis of a Single Congested Node," *IEEE ICNP*, October 2004.

[9] X. Liu, K. Ravindran, B. Liu, and D. Loguinov, "Single-Hop Probing Asymptotics in Available Bandwidth Estimation: Sample-Path Analysis," *ACM IMC*, October 2004.

[10] X. Liu, K. Ravindran, and D. Loguinov, "Multi-Hop Probing Asymptotics in Available Bandwidth Estimation: Stochastic Analysis," Technical report, CUNY, Available at http://www.cs.gc.cuny.edu/tr/TR-2005010.pdf, August 2005.

[11] X. Liu, K. Ravindran, and D. Loguinov, "What Signals Do Packet-pair Dispersions Carry?," *IEEE INFOCOM*, March 2005.

[12] W. Matragi, K. Sohraby, and C. Bisdikian, "Jitter Calculus in ATM Networks: Multiple Nodes," *IEEE/ACM Transctions on Networking*, 5(1):122–133, 1997.

[13] B. Melander, M. Bjorkman, and P. Gunningberg, "A New End-to-End Probing and Analysis Method for Estimating Bandwidth Bottlenecks," *IEEE Globecom Global Internet Symposium*, November 2000.

[14] B. Melander, M. Bjorkman, and P. Gunningberg, "Regression-Based Available Bandwidth Measurements," *SPECTS*, July 2002.

[15] Y. Ohba, M. Murata, and H. Miyahara, "Analysis of Interdeparture Processes for Bursty Traffic in ATM Networks," *IEEE Journal on Selected Areas in Communications*, 9, 1991.

[16] V. Ribeiro, R. Riedi, R. Baraniuk, J. Navratil, and L. Cottrell, "pathChirp: Efficient Available Bandwidth Estimation for Network Paths," *Passive and Active Measurement Workshop*, 2003.

[17] J. Strauss, D. Katabi, and F. Kaashoek, "A measurement study of available bandwidth estimation tools," *ACM IMC*, 2003.

[18] W. Szczotka, "Stationary representation of queues. I.," *Advance in Applied Probability*, 18:815–848, 1986.

[19] W. Szczotka, "Stationary representation of queues. II.," *Advance in Applied Probability*, 18:849–859, 1986.

[20] R. Wolff. *Stochastic modeling and the theory of queues*. Prentice hall, 1989.

Notes

[1] In general, the tight link can be different from the link with the minimum capacity, which we refer to as the *narrow* link of $\mathcal{P}$.

[2] We use the term "fluid" and "constant-rate fluid" interchangeably.

[3] The analysis assumes infinite buffer space at each router.

[4] The term Ω_i represents the volume of fluid cross-traffic buffered between the packet-pair in the outgoing queue of link L_i. For an analogical understanding, we can view the packet-pair as a bus, the cross-traffic as passengers, and the routers as bus stations. Then, Ω_i is the amount of cross-traffic picked up by the packet-pair at link L_i as well as all the upstream links of L_i. This cross-traffic will traverse over link L_i due to the flows' routing decision.

[5] Note that the turning points in $\mathcal{F}$ is indexed according to the decreasing order of their values. The reason will be clear shortly when we discuss the rate response curve.

[6] Note that the hop available bandwidth of link L_i that is of measurement interest, given by $A_i = C_i - \mathbf{x}\mathbf{r}_i$ can be less than $C_i - \mathbf{x}\mathbf{p}$.

[7] Note that the output dispersion process can be correlated. However, this does not affect the sample-path time average of the process.

[8] See section 3.2 in [9] for more discussions about this term in a single-hop context, where R_i is referred to as *intrusion residual*.

[9] Refer to [20, pages 89] for the definition of regenerative processes.