On the Scope of Active Probing Methods

Attila Pásztor\textsuperscript{1,2}, Darryl Veitch\textsuperscript{1}

Abstract—We provide a firm foundation for the estimation of route and traffic metrics using active probing. The central concepts of Initial and Busy probes are identified then used to define four signature types. Precise conditions for signatures to survive end-to-end are given, and illustrated with simulation. Techniques are provided for the efficient amplification, extraction, and interpretation of signatures, notably peaks, leading to canonical estimation methods, some of them new, for link rate, cross traffic, and available bandwidth. The concept of secondary bottlenecks is clearly defined, and a symmetry law for delay variation given. We show how all link rates can be measured, not only bottlenecks. The effects of independently varying probe size and inter-probe separation are explained. Finally packet size dependencies of link rate estimates observed in high precision measurements are explained.

Keywords—Active Measurement, delay variation, bottleneck bandwidth, internet measurement, available bandwidth.

I. INTRODUCTION

The importance of active measurement is growing rapidly, largely because of its ease of deployment, enormous flexibility, and natural relationship to the important end-to-end metrics of loss and delay in the Internet. Active probing refers to injecting controlled test traffic into a network, and measuring at a receiving node or nodes. Existing large scale active measurement programs such as Surveyor and Ripe [1] have used active probing to measure connectivity, delay, and loss statistics on coarse time scales, seconds or more commonly, minutes. Typically simple probe designs have been used such as isolated probes, and low rate periodic or (pseudo) Poisson streams. On finer time scales, several milliseconds to seconds, periodic as well as more sophisticated probe streams have been employed to measure bottleneck bandwidth [2], [3], [4], [5], [6], [7] available bandwidth [3], [6] and cross traffic [8], and to investigate the detailed statistical structure of delay and loss [3], [9], [10] of end-to-end paths.

Despite much recent activity, the full potential of probing techniques, including as they do a wealth of design possibilities across the dimensions of packet type, size, and departure process, has only just begun to be realised. More fundamentally however, it has not been clear even in principle what the scope of such methods is. Typically, techniques reported in the literature revolve around a single heuristic idea, often using a single hop model of the route, with relatively little attention paid to when the underlying assumptions used will apply. Connections between different techniques, and how they could be generalised or improved, have remained somewhat obscure. In this paper we substantially address these issues at several levels. A unifying foundation is presented with clear principles, enabling current methods to be critically assessed and put into perspective, the scope of active probing to be seen and understood, and new methods to be developed with much greater insight than before. We take our inspiration from the seminal modelling work of Bolot [2], and the practical emphasis of Paxson [3]. Technically our work is closest to Dovrolis et.al [7] through the emphasis on histograms, although the approach here is more systematic as well as broader, and we offer several new insights for probe stream design, and measurement analysis.

We construct the foundations by providing an exhaustive sample path based queueing analysis of the underlying phenomena capable of generating useful signatures of route structure and conditions. Four signature types are identified: independence, accumulation, rate and distribution. We show how they arise, and then survive or alternatively are erased by the passage through subsequent hops. Analysis techniques are developed, including histogram decomposition and ultra high resolution histograms, which enable signatures to be preferentially selected, amplified, interpreted, and then used as the basis for measurement. Eight different canonical measurement methods are provided for link rate, bottleneck rate, available bandwidth, and cross traffic estimation. Design principles are given for the probe stream which reveal the connections between probe rate and signature strength, an essential point to reduce the invasiveness of active measurement, and the separate roles of probe size and inter-probe separation are detailed. Throughout, an intuitive description is developed alongside the more formal one, which allows the phenomena to be visually grasped. Finally, real measurements are made using a high precision measurement infrastructure described in [11], and unexpected packet size dependencies are observed and explained.

It is beyond the scope of this paper to compare and contrast existing probing techniques with those developed...
here. The reader is referred to [12], [6], [7] for readable recent surveys. In section 4 we briefly indicate the underlying connections with some recent methods.

The raw timing data of an active measurement is the end-to-end delays experienced by probes. In this paper we focus on the difference of delay, delay variation, and the inter-arrival times of probes, but not delay itself. This is because neither require global clock synchronisation, but only that the clock rates are sufficiently accurate over inter-probe interval time-scales. This means that, as described in detail in [13], if a suitable software clock is used, measurements with errors well under a millisecond are possible without the need for additional hardware, such as the GPS synchronised DAG cards of the reference system we use here (see http://dag.cs.waikato.ac.nz/).

Before beginning, we invite the reader to ponder the ‘visible queueing’ of figure 1, a measurement made over a 13 hop route between the EMULab in Melbourne and the WAND group in Waikato New Zealand. The top plot shows a probe rate above the bottleneck rate, generating increasing delay lines which we call unstable bunching of probes, whereas the lower probe rate in the lower plot exhibits decreasing delay lines, stable bunching. This plot motivated our work. The sub-millisecond accuracy allows the delay lines to be traced down to small scales, and we asked, how can we collect and combine these obviously related signatures? This is one of the questions we resolve below.

Fig. 1. Delay measurement of a probe stream with a transmission rate exceeding the bottleneck bandwidth (2Mbps stream) compared to a stream remaining below it (0.3Mbps).

II. FUNDAMENTALS OF ACTIVE PROBING

We wish to define and analyse an active probing experiment, where a probe stream is sent across a network route from a source to a receiver monitoring point, which may be co-located (round-trip end-to-end), or separated (one-way end-to-end). Almost all of what we develop below holds true for passive end-to-end measurement also, where strict non-invasiveness is an additional advantage. However, in that case the probe stream is not controlled, as it is simply a subset of packets that are observed at each end, and so may be inappropriate to the needs of measurement, for example they may be absent!

A. Elements of a Probing Experiment

A route is modeled as a series of \( H \) hops in tandem. Each hop consists of a FIFO queue with buffer size \( b^h \), and an output link of rate \( \mu^h \) and transmission delay \( D^h \), \( h = 1, 2 \ldots H \). Inside a hop, packets arrive to the queue instantaneously after traversing the previous hop, and the deterministic service has the same rate as the output link: packets simply flow out. A probe stream is a numbered sequence of \( n \) packets, their protocol types, sizes \( p_i \), and departure times \( DT_i, i = 1, 2 \ldots n \). The outcome of the experiment is the arrival times \( AT_i \) of each packet reaching the receiver and a binary vector \( L_i \) indicating loss, together with the input stream specification. From these the delay and other key series are readily calculated. The metrics to measure are the physical route parameters: \( H, b^h, \mu^h \), and in particular the bottleneck bandwidth \( \mu^B = \min(\mu^h) \), the route conditions: delay, loss, and available bandwidth, and the route environment: the cross or background traffic, i.e. non-probe traffic, which enters and/or exits at each hop.

Although very rich, the above necessarily neglects a number of behaviours: scheduling is not necessarily FIFO (indeed probing could be used to determine the scheduling used), and route changes and links split at the transmission level are possible [14]. We also ignore packet reordering and fragmentation: such effects can be filtered out.

Our focus is on delay measurement using UDP/IP probes. Loss is neglected unless otherwise stated, thus buffers are assumed infinite. We aim to fully exploit both delay variation and inter-arrival times to measure all link rates, available bandwidth, and cross traffic. The use of the IP Time To Live (TTL) field to control where packets drop out along the route will be explored briefly. The route length \( H \) can easily be obtained using TTL (for example using traceroute [14]) and will not be discussed further.

A Note on Router Architecture The hop model is consistent with a store and forward switch fabric where packet arrivals and departures are counted from the end of the last bit: the input queue is seen as part of the previous transmission link, and only when the packet has fully arrived is it transferred (in negligible time) to the output queue, which is the queue modelled. After queuing the packets are serialized onto the link at link speed, and leave the queue with the last bit. With this convention the link propagation time is independent of packet size.

Concerning PASTA The above system is in some sense ‘just’ a queueing problem. However, it is very different to the traditional problems in queueing networks. Formerly, equilibrium statistics at fixed points in space (the queues) were central. In contrast, here not only is the system observed (and perturbed) by what is generally a non-typical and even non-stationary set of observers (the probes), but they each experience a different slice of the space-time environment: the route. In particular, the Poisson Arrivals See Time Averages or PASTA result of simple queueing systems [15], which has been used to justify the use of (pseudo) Poisson probe streams, does not apply in any way in this environment, observations of a route state variable are not being made!
B. The Anatomy of a Single Hop

In this section we examine exhaustively the fundamental effects due to a single hop. In a sense we wish to understand a hop as an operator, which transforms an input probe stream to an output probe stream. A route could then be seen as a composition of such operators. The approach is firmly sample path based, rather than probabilistic, and is therefore entirely general. To fix basic concepts first a general stream of packet arrivals is considered, and then probe and cross traffic packets are distinguished.

The \( i \)th packet arrives to the hop by appearing in the queue at the time instant \( \tau_i \). It begins service after a waiting time of \( w_i \geq 0 \), completes it after a service time of \( x_i > 0 \), and after a constant transmission delay of \( D > 0 \), exits the hop at time \( \tau_i^* \). The delay is therefore

\[
d_i \equiv \tau_i^* - \tau_i = w_i + x_i + D,
\]

and comparing two adjacent packets* we have the

\begin{align*}
\text{inter-arrival time:} \quad t_i &\equiv \tau_i - \tau_{i-1}, \\
\text{inter-departure time:} \quad t_i^* &\equiv \tau_i^* - \tau_{i-1}^*, \\
\text{delay variation:} \quad \delta_i &\equiv d_i - d_{i-1} = t_i^* - t_i
\end{align*}

in which \( D \) plays no role. Note that the departure time at one hop is just the arrival time at the next.

Fig. 2. System Busy & Idle periods, and \( b \) & \( i \) packets.

Lindley’s equation connects the waiting times of successive packets. With an infinite buffer, it states

\[
w_i = [w_{i-1} + x_{i-1} - t_i]^+ \quad (6)
\]

where \([x]^+ = \max(0, x)\). As illustrated in figure 2, time can be partitioned into strictly alternating busy periods where the queue is never empty, and idle periods where it is empty. By definition, the first packet of a busy period arrives to an empty queue. The non-linearity \([\cdot]^+\) of (6) therefore impacts on such idle packets. However, all other packets in the busy period experience a linear system where \( w_i = w_{i-1} + x_{i-1} - t_i \). The different ‘physics’ experienced by the idle and busy type packets is a logical yet powerful distinction which forms the backbone of the discussion below.

B.1 Relating Adjacent Packets

The expressions for the delay variation \( \delta_i \) and inter-departure time \( t_i^* \) for packet \( i \), according to its type, are

\[
i : w_i = 0, \quad \left\{ \begin{array}{l} \delta_i = (x_i - x_{i-1}) - w_{i-1} \\
t_i^* = \delta_i + t_i \\
w_i \geq 0, \quad \left\{ \begin{array}{l} \delta_i = x_i - t_i \\
t_i^* = x_i. \end{array} \right. \end{array} \right.
\]

Consider case \( i \). As \( w_i = 0 \) the previous packet does not affect packet \( i \), however as \( \delta_i \) and \( t_i^* \) compare packets, its history nonetheless plays a role. We can distinguish two subcases, depending as packet \( i - 1 \) was itself of \( i \) or \( b \) type, called \( ii \) and \( bi \) respectively (the symbols are in time order, see figure 2). Recall that positive \( \delta_i \) corresponds to increasing delay.

Case \( ii \) This is the extreme where \( w_{i-1} = 0 \) also: each packet sees an empty system and \( \delta_i \) is determined solely by the difference of the \( x_i \) series (essentially the packet sizes). Both positive and negative delay variation is clearly possible. For \( t_i^* \) the modulation of the inter-departure series, again an input one can choose, combines with the \( x_i \) series.

Case \( bi \) Here \( w_{i-1} > 0 \), which acts to decrease \( \delta_i \). If it is small, the discussion for \( ii \) holds true, alternatively \( w_{i-1} \) can dominate and \( -\delta_i \) can be arbitrarily large (subject to the constraint \( w_{i-1} < t_i - x_{i-1} \), which defines \( bi \)). Regardless of \( w_{i-1} \), if \( x_i < x_{i-1} \), delay variation will be negative.

Now consider case \( b \). The key fact is that packets \( i \) and \( i - 1 \) are back-to-back (BtB), so packet \( i \) is now directly affected by \( i - 1 \). However despite this, the history \( w_{i-1} \) of \( i - 1 \) is irrelevant for the differences \( \delta_i \) and \( t_i^* \), as the fact that the packets are BtB implies that they subsequently share that history. Packet \( i \) has ‘caught up’ to packet \( i - 1 \) as much as possible, that is by a time interval \( t_i - x_i \) which is simply their initial inter-arrival separation, minus the physical barrier due to the size of packet \( i \). The delay variation can take any real value. The interarrival time is just \( x_i \), this is the well known ‘spacing’ effect due to packets exiting the link BtB.

Again there are two subcases, \( bb \) and \( ib \), which are very similar since \( w_{i-1} \) is largely irrelevant for \( b \). Simply, since packet \( i - 1 \) is idle, the queue is likely to be smaller in \( ib \).

B.2 Relating Successive Probes

We now consider that we are restricted to measuring only some of the packets, the probes. As seen in figure 3, we introduce the notion of Initial and Busy types for probes, whilst retaining system based idle and busy periods. Initial probes are again the first (of their type) in a busy period, but they may not be the first packets. Other probes in the same busy period are of \( B \) type, and will have to queue behind some combination of probe and cross traffic (CT). The distinction is important as the queueing of \( f \) probes is decoupled from that of other probes, whereas \( B \) probes are
connected to those ahead in a fixed, linear way. Henceforth
i indexes probe packets only.

The analogues of (7), (8) are \( \{ w_i \geq 0 \text{ in both cases} \} \)

\[
I : \begin{cases} 
\delta_i = (x_i - x_{i-1}) + (w_i - w_{i-1}) \\
\tau_i = \delta_i + t_i 
\end{cases} \\
B : \begin{cases} 
\delta_i = x_i - t_i + c_i \\
\tau_i = x_i + c_i, 
\end{cases}
\tag{9}
\]

where \( c_i \) is the aggregate service time of the CT entering the
queue between probes \( i - 1 \) and \( i \). Equation (9) is in
fact true for all probes, whereas (10) is for \( B \) probes only.
To understand the consequences of the Initial vs Busy

Fig. 3. B & I probes. In time order the probes are of types: \( I, I, B, I, B, B \), and more precisely: ‘\( I, I, I, B, B, B \)’
probe distinction, consider first the set \( I \). Equations (9) can
each be decomposed into a term controlled by probe char-
acteristics, and a queueing term related only to cross traf-
fic, since preceeding probes were in earlier busy periods
(see caveat below). As before \( \delta_i \) can take either sign, and
a hard per-packet lower limit of \( \delta_i = x_i - t_i \) occurs when
\( w_i = 0 \) (minimum wait), and \( w_{i-1} = t_i - x_{i-1} \) (maximum
wait consistent with \( I \), implying \( c_i = 0 \)). For \( \tau_i \) this corre-
sponds to a lower limit of \( x_i \), that is ‘almost BtB’ output.
Another constraint is that \( c_i < x_{i-1} - t_i \), the CT must not
‘fill the gap’. The set \( B \) of busy probes follows different
laws, equations (10), but which also feature separate probe
and CT terms. Just as for \( B \) packets \( \delta_i \) can take any real
value, but locally CT tends to increase it, and to increase \( \tau_i \)
by forcing apart packets on the link.

Rather than organising the sequel about the subsets \( II \)
and \( BI \) of \( I \), and \( IB \) and \( BB \) of \( B \), it is instructive to intro-
duce the extreme cases of an independent system, where
all probes are \( I \), that is there is at most one in any busy per-
iod, and a linear system, where all probes are \( BB \) (except
the first), that is they are all in the same busy period. To as-
sist in this, let us adopt a compact descriptive language for
certain key mechanisms, or ‘forces’ acting on probes. By
probe separation we mean the increase in inter-probe time
that can result from queueing in general and cross traffic in
particular, and by bunching we refer to probes becom-
ing BtB as a result of queueing. These two relate closely to
delay variation. The third relates to inter-departure time, it
is the well known spacing effect where the link rate sets
the separation of BtB probes, generating a rate signature
in the \( \tau_i \) series.

In what follows randomness is sometimes introduced,
for example to avoid periodic CT, however it is important
to understand that the resulting sample paths have been
post-selected to be independent systems or otherwise as
required – the approach remains sample path based.

B.3 Independent System Behaviour

Figure 4a shows the delay histogram from a
simulation of a 1 hop independent system probed by a 1p-
kt/s periodic stream of 1500byte packets, with 0.55Mbps
Poisson cross traffic with packet sizes uniformly distrib-
uted over \([0,250]\), and \( \mu = 0.75\)Mbps. The wide spacing
of the probes makes an independent system very likely,
so finding sample paths which are independent systems is
easy. The clear peak at zero corresponds to packets with
zero waiting time, and the spread about it might be inter-
preted as noise due to the cross traffic term, since the pack-
et size component cancels. Plots like these are familiar from
low resolution studies such as those of [16], however
the conditions under which it is valid to consider the CT as
noise are typically ignored. Here we emphasize the need
for an independent system, and note that, if a probabilistic
model was selected for \( w_i \), it would be largely independent
of the \( x_i \), and furthermore as diff acts to reduce correla-
tion, an i.i.d. model for \( \text{diff}(w_i) \) is in fact not unreasonable.

An important point is that \( w_i \) cannot be entirely inde-
dependent from the probe characteristics for two reasons,

i) the probes do not observe the system at stationary times,

ii) the fact that we have selected the sample path to be
an independent system places a subtle conditioning on the
allowed \( w_i \). We return to this point repeatedly.

Fig. 4. Independent System. Effect of probe size distribution
on delay variation. Simulation: (a) \( p_t = p \), (b) alternating,
(c) uniformly distributed \( p_t \). Measurement: (d) alternating.

In figure 4(b) probe sizes were made to alternate be-
tween 1500 and 40bytes, inducing an alternating \( x_i \)
series. The histogram can be interpreted as the distribution of
\( \text{diff}(x_i) \), with a noise which is \( \text{diff}(w_i) \). By increasing the
number of different packet sizes richer histograms result,
for example in figure 4(c) they were uniformly distributed
between 0 and 1500 bytes. Such continuity disguises the
separate nature of the probe and cross-traffic effects.

In independent systems (but not for \( BB \) probes in general)
\( \delta_i \) also has an upper bound, namely \( t_{i+1} - x_{i-1} \), express-
ing the coordination between neighboring hops required to
keep all probes \( I \). We also have an approximate sample
path symmetry law for delay variation deriving from
\( \sum_{i=2}^{n} \delta_i = (x_n - x_1) + (w_n - w_1) \), which expresses the
fact that a delay variation of one packet affects its neigh-
bours’ variations by equal and opposite amounts, generat-
ing a symmetry that is only broken by the edge effects, the
values at \( i = 2, n \). This is very apparent in figure 4, and
can be used to disprove an independent system hypothesis,
as well as test for the presence of loss, both of which will
break the symmetry.
B.4 Linear System Behaviour

A linear system is inherently unstable or marginally stable, that is its average input rate is at least \( \mu \). For example if \( x_i = x, t_i = t \) and \( c_i = c \) are all constant then delay will increase steadily with rate \( \delta = x - t + c \geq 0 \), yielding a single peak in both the \( \delta_i \) and \( t_i^* \) histograms. Distinct multiple peaks can clearly be obtained by using a small number of different packet sizes and separations. We now introduce four key phenomena intrinsic to linear systems.

Take a Poisson cross traffic sample path with uniformly distributed packet size, and select a periodic probe stream with constant \( x_i, t_i \), such that \( x < t \) and the system is indeed linear. Thus the average probe arrival rate \( x/t \) is below \( \mu \), but combined with CT it is sufficient to produce a linear system. The corresponding delay variation is shown in figure 5(a). The peak at the minimum \( \delta_i \) illustrates the first key phenomenon: thanks to CT building up the queue, bunching occurs allowing successive probes to experience a steady decline in delay as the CT is served. We refer to this as stable bunching, as it occurs despite the probe rate being under \( \mu \), and generates a peak at negative \( \delta_i \).

Next, allow the same CT to multiplex with probes with the same \( t \), but with \( x > t \). Figure 5(b) shows the second phenomenon: unstable bunching, which generates a peak at positive \( \delta_i \) corresponding to the queue building due to the unstable probe rate, without the help of CT. Of equal importance is the third effect: when increasing the probe rate via \( x \), the \( \delta_i \) histogram remains (strictly) invariant.

Now we achieve the same increase in rate, not by increasing \( x \), but by decreasing the inter-probe separation \( t \). Figure 5(c) shows two effects, that the minimum \( \delta_i \) now takes a different value (see equation (10)), and the fourth key result: increasing probe rate via \( t \) or \( x \) are not equivalent, different distributions result. This is explained by the fact that, for a given CT, changing \( t \) changes the distribution of the number of packets falling between the probes.

Since \( t_i^* = \delta_i + t_i \), at constant \( t_i \) the inter-arrival time histograms are just those of the top row in figure 5 shifted to the right by \( t \). Rather than showing these, in the lower row we give \( t_i^* \) for the same experiments except that the \( t_i \) are not constant but Poisson distributed (with means as before). In general changing \( t_i \) changes the histogram shape, however, provided the probe size is constant the peaks remain sharp, since when probes bunch the details of the CT become irrelevant. On the other hand variable \( t_i \) ’s produce a variable shift in the \( \delta_i \) plots even with fixed \( x_i \), broadening the peaks (not shown).

By considering a probabilistic model of how many CT packets fall between probe packets, the histogram shapes in figure 5 can be explained by using approximate convolution arguments. For example the rectangular shape of figures 5(c) and (f) correspond to a single CT packet of uniformly distributed size, the dominant scenario due to the smaller average \( t \) value. The more complex shapes in the other plots relate to the dominance of one or two packets, whereas the peaks in all cases are the signature of zero packets. When the number of packets becomes large the distributions become smooth and bell shaped.

B.5 Mixed Systems, and Probe Stream Decomposition

The independent and linear systems are useful extremes, but seem restrictive, especially the latter which implies a highly loaded system and/or highly invasive probing. One of the major contributions of this paper is to demonstrate that this is not the case: the probe stream can be seen as a patchwork of short independent and linear systems, connected by transitions. They can be isolated by decomposing \( \delta_i \) and \( t_i^* \) histograms into sub-histograms for the probe sets \( I, B \), or more finely, \( XI, XB, X \in \{I, B\} \), and analysed as before. The simplicity of the simple systems can therefore be meaningfully exploited in general cases. We construct a more ambitious thought experiment, beginning with a linear system and decreasing the probe rate until an independent system is reached. Again \( x_i = x \) and \( t_i = t \) within each plot for simplicity, and a fixed CT sample path with uniformly varying packet sizes is used. In the top row of figure 6 we recognise the independent system \( \delta_i \) histogram on the left and the linear on the right. In the mixed case (c) with smaller \( t \) we see that the single peak of plot (d) has split, separates further in (b), the rightmost becoming the central peak of the independent system in (a), the other dissipating its mass out to the left. In the bottom row from left to right, we decompose plot (c) into \( II, BI, IB \) and \( BB \) histograms, that is we simply take the \( \delta_i \) values separately according to probe type. The full histogram (assuming the same bins) is literally the sum of these components. A similar decomposition can be performed for \( t_i^* \). The \( II \) and \( BB \) components exhibit the features of an independent and linear system respectively: we have succeeded in isolating them from the mixed system. Note that the \( II \) component is nonetheless visibly subject to the constraint on its minimum noted earlier. The decomposition of plot (b) is similar, we just see a natural shift in mass to the \( II \) histogram, as the independent system is approached.

The transitional \( IB \) and \( BI \) cases are more complex. Although \( IB \) is governed by the same equation as \( BB \), and
BLI by the same as II, there are important differences as seen in figure 6. These can be explained in terms of the conditioning of the waiting times of packet $i - 1$, leading to negative $\delta_i$ being less common for BI than for BB, but more common for BL than for II. Roughly speaking, mass in the XB histogram to the left of zero is transferred preferentially to BI and then on to II, as probes progressively change type with reducing rate.

Fig. 6. Transition through systems. Top row: $\delta_i$ with steadily increasing probe rate beginning with an idle system (a), passing through mixed systems: (b), (c), to linear system: (d). Bottom: histogram decomposition for system (c).

III. A NONLINEAR SURVIVAL COURSE: N HOPS

If the previous section described the hop ‘operator’, the present one tackles the issue of operator composition. The superscript $h$ indicates the hop, quantities without them pertain to the route. The defining relations across hops are

inter-arrival time operator: $t^h_i = t^{h-1}_i + \delta^h_i$ (11)

route delay variation: $\delta_i \equiv \sum_{h=1}^{H} \delta^h_i$ (12)

route inter-departure time: $t^*_i = t_i + \delta_i$. (13)

A. The Second Hop

Each probe $i$ is either I or B at each hop, yielding four combinations over 2 hops: I-I, B-I, I-B and B-B. Combining equations (11)–(13) together with (9) and (10) yields

I-I \( \delta_i = \sum_{h=1}^{2} \left( (x^h_i - x_{i-1}^h) + (w^h_i - w_{i-1}^h) \right) \) (14)

\[ t^*_i = \delta_i + t_i \]

B-I \( \delta_i = (x^2_i - t_1^i + c^2_i) + (x^2_i - x_{i-1}^2) + (w^2_i - w_{i-1}^2) \)

\[ t^*_i = (x^2_i + c^2_i) + (x^2_i - x_{i-1}^2) + (w^2_i - w_{i-1}^2) \] (15)

I-B \( \delta_i = x^2_i - t_i + c^2_i \)

B-B \( t^*_i = x^2_i + c^2_i \) (16)

as the complete set of defining equations, which we examine in terms of how the delay variation and inter-departure time signatures of probes exiting hop 1 are modified by the passage through hop 2. Two main effects emerge: if a probe is of type I at hop 2, then (taking $x^2_i = x_{i-1}^2$ for simplicity) a noise $(w^2_i - w_{i-1}^2)$ is added, whereas if is of type B, then the hop 1 signature is erased and replaced with a B type signature corresponding to the route input $t_i$ arriving to hop 2.

We add two caveats to this first order analysis. First, the ‘noise’ in fact contains correlations which may be strong, especially if a lot of cross traffic follows the same route as the probes. Second, according to equation (16) I-B and B-B are formally equivalent to a single hop, however, $c_i^2$ enters hop 2 during a time interval of $t^*_i$, not $t_i = t^1_i$, which can change its distribution significantly. The two are equal only if $\delta^1_i = 0$, which is more valid for I-B. These second order considerations relate to the finer II, BI, IB, BB distinctions of each probe at each hop.

B. Running the Gauntlet: N Hops

The observations of the previous section motivate the pursuit of an informed heuristic approach to the N hop problem, avoiding the combinatoric explosion of a complete, direct analysis.

The Survival Heuristic

Consider probe $i$ which acquires a signature at hop $h = s$. This signature will survive to the receiver if at all subsequent hops probe $i$ is of type I (survival will be stronger when probes are in fact II, due to the symmetry of IIDependent systems). If route conditions are such that enough such probes survive, then the signature may be detected at the receiver in the $\delta_i$ and/or $t^*_i$ series.

Two main classes of survival are possible, when the probe is B at some hop $s$ and I afterward, or when it is I over all hops. The latter case is particularly simple and is illustrated in a 13 hop measurement in figure 4(d), where we recognise the characteristic symmetric independence signature of an IIDependent system histogram (each probe being in fact II), created at hop 1, and then surviving through downstream noise. However, in figure 4(d) a probe size $p_i$ dependent term is also accumulated at each hop, as described by the generalisation of equation (14) with $x^h_i = p_i/\mu^h$:

\[ \delta_i = \left( p_i - p_{i-1} \right) \sum_{h=1}^{H} \frac{1}{\mu^h} + \sum_{h=1}^{H} (w^h_i - w_{i-1}^h). \] (17)

Such an accumulation signature is generated by any sequence of hops at which the probe is I, it is independent of CT, and arises whenever probe sizes are unequal.

A key question is whether information can be extracted from more than one hop. We now show that signatures can in principle be obtained, depending on circumstances, from all hops, through the survival of B probes. The governing equation, generalising equations (16), (15), is

\[ \delta_i = (x^s_i - t_i + c^s_i) + (p_i - p_{i-1}) \sum_{h=s+1}^{H} \frac{1}{\mu^h} + \sum_{h=s+1}^{H} (w^h_i - w_{i-1}^h). \] (18)

where $s$ is the last hop at which probe $i$ was busy. For simplicity however we often set $p_i = p$. 
B.1 Rate Signatures, Secondary Bottlenecks

Consider a rate signature at hop \( s \), that is the spacing of BtB packets traditionally used to detect the bottleneck (BN) rate \( \mu^B \). We already know that rate signatures can in fact be created at any hop, via stable or unstable bunching, depending on cross traffic conditions. As for the survival of such a signature, the situation is exactly as for the bottleneck link: if \( \mu_i > \mu_s \) for all downstream hops \( h \in [s + 1, H] \), then it can survive, again depending on the CT. This motivates the following concept.

Definition: Secondary Bottlenecks (SB) are those hops downstream of the bottleneck for which no hop further downstream has a higher rate.

Two general properties of SB’s are:

- the rates of SB’s monotonically increase with hop.
- the probe rate is throttled by the bottleneck, thus the rate signature of all SB’s are due to stable bunching, the peaks must be at negative \( \delta_i \) (the BN can have either sign).

In figure 7, a simulation of a three hop route with rates \( \{\mu^h\} = \{10, 2, 100\} \text{Mbps} \) and CT of \( \{2, 1.5, 50\} \text{Mbps} \) is given, with a probe stream of average rate only 6.36kbps. The stream is packet pair with 0.1s inter-pair spacing, and in plot (a) where \( p = 40 \text{bytes} \) the pair spacing is 0.64ms (0.5Mbps local rate). The 2Mbps bottleneck rate signature is the central peak in the \( \delta_i \) sub-histogram, where probes that are B (at any hop) have been focused on through a strategy of extracting the second probe from each pair, for which \( t_i \) is constant. The small peak on the left corresponds to the 100Mbps SB, and appears at negative \( \delta_i \) as expected. A rate signature at the first hop exists but has been erased by the BN. The peak at the origin is due to the fact that the second probe of each pair is not necessarily B, thus we also illustrate here how informed guesses of probe type can be used to guide decomposition, and how imperfect selections may nonetheless be usefully interpreted. In plot (b) a similar experiment was performed with \( p = 1500 \text{bytes} \), with the same probe rate and CT. The rate signatures are much weaker, because the larger packet gives CT more time (store & forward mechanism) to slot between the pairs.

![Fig. 7. Rate Signature Survival. Peaks in the \( \delta_i \) sub-histogram (2nd probe of each packet pair) (a) \( 40 \text{byte} \) probes: BN peak is striking, SB peak clearly visible (left) and the I probe peak (right) is small. With resolution increase of \( 10 \times \) (black), the peak/background ratio increases dramatically. (b) \( 1500 \text{byte} \) probes: I peak and noise dominates.](image)

Peaks versus Modes  The black histogram overlayed in figure 7(a) is the same data plotted with a bin size of only \( 5 \mu s \), compared to \( 50 \mu s \) for the grey. The improvement in ‘signal to noise ratio’ is dramatic: the peaks are weakly affected, whereas the values in other bins have dropped significantly. The use of such ultra high resolution histograms (UHRH) is justified by the inherent high speed of networks. Those BtB probes which survive with no or little downstream queueing will have a very precisely defined \( t_i^* \) (provided the measurement infrastructure is adequate [11]). Their detection with UHRH’s is a powerful technique which is quite different from the idea of detecting ‘modes’, which is related to general histogram shape and is a complex mix of many factors.

Probe Design  For the simultaneous measurement of multiple rate signatures it is advantageous to send out probes with local rates (close probe pairs, triples..) as high as possible to encourage bunching. Thus it is important that the sender have a high access speed. As explained above, small packet sizes also encourage bunching by reducing the probability of CT separation.

B.2 Looking Upstream of the Bottleneck

In general only certain hops, and in particular none upstream of the BN, can be SB’s, which seems to indicate that they are invisible to probes and hence unmeasurable. However, SB’s are based on rate signature, which is not the only kind of signature possible even for B probes. The key is that probes can be B without being back to back if they experience CT separation. Indeed, to satisfy the necessary survival condition that a probe be B at the hop \( s \) of interest and I subsequently, even before the bottleneck(s), the separation must be such that the local departure rate of probes at \( s \) is lower than all downstream link rates, to avoid obligatory erasure. Such conditions can not be satisfied at just any hop. For example in the packet pair probing of the 3 hop route of figure 7, to generate a signature at the first hop which can pass the bottlenecks we require a packet pair separation corresponding to a local rate below 2Mbps, yet a cross traffic capable of making sufficient probes type B at hop 1. Whether this is possible depends on the CT, however in general it is possible at one hop at least, that with the smallest available bandwidth (AB). In a sample path description AB is a concept relative to what individual probes see, however roughly speaking it refers to the difference between \( \mu^h \) and CT rate \( \lambda^h \). The signature in such a case is the entire distribution of \( c_i^h \), the CT that enters hop \( s \) over the time interval \( t_i^* \), an example of which is given in figure 8 for cross traffic rates of \( \{9,8,0.5,50\} \). We describe how to make use of such a distribution signature in the next section.

Probe Design  The requirements of distribution signatures are opposite to those of rate signatures. It is essential to avoid BtB probes and associated peaks, so large packet
sizes are appropriate. To find the signature one can begin with a large probe pair separation and decrease it until the symmetry of the sub-histogram (2nd packet of pair as above) is broken, indicating $B$ probes, which will be generated in the majority at the hop with the smallest AB, by definition (this could be at any hop, even the BN).

**Using the TTL Field** Cross traffic plays a very important role in signature generation by inducing both stable bunching and widely separated $B$ probes, and as such which signatures are measurable will be time varying. One approach is to perform measurements at different times to benefit from a variety of route environments, another is to assist the CT by injecting other streams. The TTL field can be used to measure delays to hops midstream directly, however the timestamps are unreliable due to the way routers treat ICMP packets. However, it is ideally suited to keeping queues non-empty at a target router with packets which do not proceed further, limiting the invasiveness of the assisted CT procedure downstream.

### IV. THE BASES FOR ESTIMATION

Many of the effects above can be exploited for measurement. Below and in table IV-C we catalogue the canonical approaches corresponding to the four signature types of a H-hop route. ‘Equivalent 1-hop’ models serve at best to model independence signatures, and are of limited use. To emphasize this, note that it is not even possible to have an ‘invisible’ hop $h$ (identity hop operator): even if CT was zero with $\mu^h = \mu^{h-1}$, modulating probe size would reveal it to the $\delta_i$ series, it would have an accumulation signature. Hybrid approaches, and optimisation of the techniques, is a subject of ongoing work.

#### A. Rate Estimation

**R1: I Probe Modulation** The method is based on independence and accumulation signatures coexisting as per equation (17). From a $\delta_i$ histogram such as figure 4(d), where alternating packet sizes $p_1, p_2$ create a bimodal distribution about symmetric peaks at $\delta_i = \pm \Delta$, a statistic involving all link rates can be accessed via

$$\sum_{h=1}^{H} \frac{1}{\mu^h} = \frac{2\Delta}{|p_1 - p_2|}. \quad (19)$$

In the example given $2\Delta = 0.0165$, corresponding to $1/\mu_j = 1.413e^{-6}$, to which the 2Mbps bottleneck contributes $0.5e^{-6}$. The distance $\Delta$ could be measured by quantiles or by moments.

**R2: Bunching** As already discussed at length in section III-B.1, equation (18) with small $p_i = p$ allows the sharp peaks of rate signatures in either of the $\delta_i$ or $t^*_i$ histograms to simultaneously reveal the BN and the SB’s.

Given such a peak, say at $t^*$, the corresponding rate is

$$\mu^s = \frac{p}{t^*}. \quad (20)$$

Note that the peak splitting effect of accumulation signatures seen in R1 could be used here, and generally, to gain information over hops downstream of $s$.

**R3: Proportional to Packet Size** From the discussion in section III-B.2, and using equation (18) with $p_i = p$, we can write $t^*_i = x^i + c^i + n(s+1, H)$, where $n(s+1, H)$ is the noise downstream. A distribution signature for $t^*_i$, such as that of the sub-histogram in figure 8(a), is then seen as the distribution of $c^i$ for the hop $s$ with the smallest AB (with added noise), with $x^i = p/\mu^s$ as an independent location parameter. Measuring the shift in the distribution as a function of $p$, for example via its mean $\overline{t^s}(p)$, one obtains $\mu^s$ by regressing on $p$ (see figure 8(b)) using

$$\overline{t^s}(p) = \text{const} + \frac{p}{\mu^s}. \quad (21)$$

Care must be taken to only vary $p$ in a range consistent with the distribution signature. This method requires that separate measurements be made at different $p$, and so relies on a kind of stationarity of the CT.

**R4: Proportional to Packet Rate** Method R4 is a variant of the approach of R3 where the $t_i$ component of probe design is exploited, as it may be undesirable to vary packet size (see below). By continuing the earlier calculations by dividing by the time $t^*_i$ over which $c^i$ enters hop $s$, we obtain $t^*_i/t^*_i = x^i/t^*_i + c^i/c^i + n(s+1, H)/t^*_i$. Changing $t^*_i$ will change the distribution of $c^i$ (and scale the noise), however $c^i/t^*_i$ represents a kind of average CT rate per unit time, which will have a constant mean assuming stationary CT. Analogously to R3, we may then extract $\mu^s$ by regressing on $1/t_i = 1/t$

$$\overline{t^s} = \text{const} + \frac{p}{\mu^s} \cdot \frac{1}{t}, \quad (22)$$

where we have substituted $t$ (the accessible probe input parameter) for $t^*_i$ on the left, via the assumption of $s$ being the link with minimum AB and hence that probes are $I$ on upstream hops, so that $t = t^*_i$.

Fig. 8. Distribution signature survival. (a) sub-$t^*_i$-histogram with no peaks for $p$ fixed. (b) mean of histogram shifts linearly with $p$ (shape is preserved), the slope gives $1/\mu^s$.

#### B. Cross Traffic Estimation

By cross traffic estimation we mean estimation of the average CT rate $\lambda^h$ entering a given hop $h$. We assume the required link rates are known, perhaps through a prior estimation phase. The main difficulty is that if $\lambda^h$ is too
small, then the queues can be empty when probes arrive, so that periods of low rate will be undersampled, perhaps severely. Some compensation is possible however through measuring the utilisation via the mass of the $\delta_i = 0$ peak.

**CT1: Queue Tracking**  At a single hop in an independent system with $p_i = p$ and $t_i = t$, $\delta_i = w_i - w_{i-1}$, which if the queue did not empty implies $\overline{\delta_i}/t = \lambda/\mu$. As in R1, independence signatures over a route arise from a weighted sum of effects across all hops: $\overline{\delta_i}/t = \sum_{h=1}^{H} \lambda^h/\mu^h$. This can be used either to calibrate a single hop conceptual model of the route, or, given partial information on some hops, to deduce bounds on the remaining ones.

**CT2: Time-scale tracking**  This method follows immediately from the discussion of R4. The $const$ in equation (22) is precisely $c^*_{i}/t^*$, the average CT rate that we would like to measure. It can be extracted as the intercept of the same regression on $1/t$, or alternatively on $p$. The same caveats apply, small rate values where the distribution signature ceases to hold will not be measured well.

C. **Available Bandwidth Estimation**

**AB1: Inferred AB**  Given any method for CT and link rate estimation on a given hop, we can define an average AB rate estimate by simply subtracting: $\mu^h = \mu^h - \chi^h$.

**AB2: Just Saturated**  Using the logic of section III-B.2 using packet pair, the lowest AB can be calculated from the measurement of $\mu^h$ and the local probe rate at which the sub-$\delta_i$-histogram loses its symmetry.

D. **A Note on Probe Size Dependence**

Thus far we have set $x_i^h = p_i/\mu_i$, however there are two important reasons why this may be inaccurate: i) packet processing: which adds a hop dependent delay $\chi^h$ independent of packet size, and ii) protocol change: which induces a per-hop header size change $\rho^h$ in each packet. Simple models for these effects are

- **packet processing:** $x_i^h = p_i/\mu_i^h + \chi^h$  \hspace{1cm} (23)
- **protocol change:** $x_i^h = (p_i + \rho^h)/\mu_i^h$  \hspace{1cm} (24)

We use these models in real experiments in the section 5.

**E. Some Relations to Previous N-Hop Methods**

Peaks from stable bunching are mentioned for example in [7], although modes rather than UHRH’s are used, and $t^*$ histograms rather than the more useful $\delta_i$. The distribution signature is related to the TOPP method [6].

V. **A Example: Packet Pair Revisited**

Two simple experiments were performed to demonstrate rate signature survival in real networks. The probe stream used is the well known Packet Pair method discussed above, however instead of trying to identify characteristic distribution modes, our analysis relies on UHRHs. The probe size dependence of the results is also explored. The experiments were performed between sites at Budapest, Melbourne and Waikato. The route from Budapest to Melbourne consists of 21 hops, with the 256 kbps Budapest access link being the BN. The Melbourne-Waikato route has 13 hops with an estimated 2Mbpks BN.

Two measurements were performed on the Budapest-Melbourne route using probe packets of 40 and 1500 bytes. The results are shown in the $t^*$ histograms of figures 9(a) and (b). The BN estimate based on the small packet measurement is 218kpbs, while the estimate with the larger probes is 255kpbs. The difference can be explained, using equation (24), as a lower layer protocol extending the IP packet size by roughly 7 bytes. The access link is in fact using HDLC which adds 6 octets. We also notice that the smaller probes clearly revealed the presence of SBs (peaks to the right of the BN have positive $\delta_i$ and cannot be SBs).

A similar measurement was performed on the Melbourne-Waikato route, with six different packet sizes logarithmically spaced between 40 and 1280 bytes. Unexpectedly, the leftmost SB peak was observed at the same value, $t^* = 0.22$ms, for the three smallest probe sizes. This can be explained by a shaper or other per packet bottleneck obeying equation (23). Returning to the BN, again the header change effect of the rate estimate is seen. The probe size dependence of the BN estimate is shown on the lower curve in figure 9(c). Only five points are plotted, because for the smallest packet size the usual BN signature was erased by the per packet BN signature. The upper line shows the estimates corrected for a 50byte additional header, the agreement is good.

**Fig. 9.** Bunching Method rate estimation in real measurements, and estimation errors due to packet size dependence. (a) small packet based UHRH showing BN and 3 SBs, (b) large packet UHRH showing only the BN signature and a larger BN rate estimate, (c) packet size dependence of BN rate estimate from a different route plus corrected estimates.
VI. CONCLUSIONS

We have provided a firm basis on which active probing methods can be described, analysed, understood, and designed. The key insight is the distinction between Initial probes, which are not directly affected by earlier probes in the stream, and Busy probes, which are. This distinction is based on the application of classical busy period ideas from queueing theory, applied across a route modelled as a tandem queue, and specialised to the observing probe stream. A sample path based approach was taken, so conclusions are not reliant on specific traffic models.

A hop was viewed as an operator on a probe stream, and a two hop system as an elementary operator composition. This led to the definition of signatures, of which four types were identified: independence: characteristic of probes which pass the entire route as I probes, seeing queueing simply as noise from cross traffic, accumulation: the inevitable packet size dependent accumulation of service time at each hop, rate: the familiar fixed spacings due to back to back probes, and finally distribution, arising from correlated probes protected from bottlenecks by cooperative cross traffic.

The Survival Heuristic was introduced which explains how the signatures can survive passage through the route once created. Histograms were highlighted as a rich means of examining both the delay variation and inter-arrival time series which carry the signatures. Histogram decompositions were introduced both to explain the separate life cycles of I and B probes, and as a powerful means of concentrating signatures by isolating the probes that carry them. The different shapes of histograms under different conditions were discussed, an important aid for their interpretation in different shapes of histograms under different conditions.

With independence signatures where I probes dominate, the natural idea of averaging away noise was formalised into the need for hops to be independent systems, and a symmetry law was introduced for delay variation.

Rate signatures are associated with B probes that bunch back to back. The important distinction between stable and unstable bunching was made allowing secondary bottlenecks, which can be detected simultaneously with the main bottleneck, to be precisely defined. It was pointed out that bottlenecks can be reliably detected through peak detection using ultra high resolution histograms, taking advantage of the high speed of networks and accurate measurement infrastructure, provided there is sufficient cross traffic. Such very sharp peaks were distinguished from the much vaguer concept of modes.

One of the main results is that although rate signatures cannot in general allow the measurement of all link rates, nonetheless this is possible, by using distribution signatures. It is explained how this is dependent on the available cross traffic.

The dependence of the histograms and signatures on the probe sizes and inter-probe separations was investigated, leading to much insight into probe design, signature enhancement, and new measurement methods. It also paves the way for a far more rigorous and informed ‘science’ of probe stream design in the future.

Finally the issue of non-linear service time of packets as a function of packet size was addressed, as an important practical issue. We illustrated and analysed two separate such effects, per-packet processing and header size dependence, in measurements using our own high precision probing infrastructure.

ACKNOWLEDGMENTS

This work was supported by Ericsson. The authors thank Rolf Riedi for helpful comments on the manuscript, and the WAND group and Ericsson Traffic Lab for measurement access.

REFERENCES


