Estimating Available Bandwidth on Access Links by Means of Stratified Probing

Bjørn J. Villa and Poul E. Heegaard
Norwegian Institute of Science and Technology, Department of Telematics, Trondheim, Norway
{bjorn.villa, poul.heegaard}@item.ntnu.no

***Abstract*—This paper presents a novel approach for estimation of available bandwidth on access links using stratified probing. The main challenges of performing such estimations in this network part is related to the bursty nature of cross-traffic and the related uncertainty regarding appropriate time period for producing sample estimates. Under the fluid flow traffic model assumption, these problems would not be present – but for the access network part this assumption does not hold. The method suggested in this paper is based on a four-phase approach. In the first phase a traffic profile for the cross-traffic is established, with focus on detecting periodic behavior and duration of respectively burst and idle sub-periods. In the second and third phase, the active probing is split into strata and synchronized according to the burst/idle sub periods. In the final phase, the actual probing and estimation of available bandwidth takes place. The method is analyzed by means of experiments in a controlled lab environment, using adaptive video streaming as the service with a periodical behavior. The empirical results are considered quite promising both in terms of accuracy and the low degree of intrusiveness facilitated by the stratified approach.**

***Index Terms*—available bandwidth estimation, stratified probing, access network, adaptive video streaming, over-the-top**

I. Introduction

The amount of services provided to Internet users around the world following an Over-The-Top service delivery model is increasing. This model is based on that the involved network operators are not taking any active measures in order to assure the quality levels of the specific services. Traffic is carried as part of the best-effort class and will therefore face obvious challenges in terms of being able to meet the end users expectations regarding Quality of Experience (QoE).

In terms of service usage, there has also been a significant change over the last decade. The dominance of services including video components is increasing. This applies both in terms of traffic volume and service usage frequency. The obvious example of such a service provider is YouTube.

For services with video components, the capacity requirement in terms of bitrate is critical as this could be in the order of several megabits per second. Up until a few years ago, delivering such services Over-The-Top was almost impossible without involving the network operators in order to gain access to other Quality of Service (QoS) classes than just best effort. As solutions for adaptive video streaming emerged and the relevant MPEG-DASH standard [1] was approved, this situation changed. By introducing additional functionality in service endpoints, these solutions make it possible to change quality level during service delivery based on certain observed parameters. Among these parameters are available bandwidth between client and server.

The specific methods for estimating available bandwidth and other interesting metrics are not part of the MPEG-DASH standardization effort. This makes it an area of particular interest for technology vendors, in their effort to make their solutions perform well in the market place. However, accurate methods for estimating available bandwidth are of interest also for other services. An interesting aspect in this regard is that the growing amount of video service on the Internet makes it even harder than before to perform available bandwidth estimations due to the embedded bursty nature in traffic generated by these services. This follows by the repeated fetch-next-video-segment operations, and associated traffic bursts from server to client.

The relevance of the research presented in this paper can be viewed as a contribution to the domain of adaptive services in general, which has the need for an indicator on how much cross-traffic is present, and subsequently available bandwidth. In this context, the term cross-traffic is used to describe the aggregate of traffic present on an access link. A continuous estimate of cross-traffic volume can be used for both adjusting quality levels on a per service basis and as input to a network qualification test prior to service usage.

A. Problem Statement

The amount of cross-traffic present on an access link during a specific interval (i), ending at time t is described in terms of number of bits sent (V_i) across it during the time interval (T_p). The available bandwidth (B_{i,T_p}) during the same interval is the difference between the total link capacity (C) and the cross-traffic estimate (cf. Eq. 1).

$$B_{i,T_p} = C - \frac{1}{T_p} V_i[t - T_p, t] \tag{1}$$

The sequence of available bandwidth estimates results in a time series B_{T_p} with index i over the index set N. The

Manuscript received June 1, 2013; revised September 3, 2013.

doi: 10.12720/ijoee.1.4.213-221

size of N corresponds to the duration of the available bandwidth estimation (cf. Eq. 2).

$$B_{T_p} = \{B_{i,T_p}, i \in N\} \quad (2)$$

Using active probing as the method to obtain estimates for the cross-traffic V_i requires a number of K probe packets (or packet pairs) to be sent and analyzed by the receiver during each time interval T_p. The results from the probing, gives a sequence of cross-traffic samples v_j indicating how many bits of cross-traffic a probe was influenced by. The cross-traffic influence is detected by measured increase in inter-arrival time for probe packets, or delayed arrival of single probe packets. In order to get the total cross-traffic estimate in number of cross-traffic bits V_i observed for the specific period i, the v_j samples are summarized (cf. Eq. 3).

$$V_i = \sum_{j=1}^{K} vj \quad (3)$$

Even though the formulas for this type of estimation are simple, providing input to them which gives an accurate result is not straightforward in cases where the cross-traffic does not follow a fluid-flow model. As will be described closer in Section II, real traffic does not follow the fluid-flow model, and in particular services with video components generate a very bursty traffic profile.

Given the bursty nature of popular services such as video streaming, and put into the context of access networks, it raises some specific challenges for performing accurate available bandwidth estimations. The problem at hand is how to perform active probing of burst oriented cross-traffic without introducing access link congestion, and also how to choose the appropriate time period T_p for computing $B_{i,Tp}$ samples.

Choosing the appropriate time interval for cross-traffic estimation can be done in different ways. From the perspective of describing the cross-traffic as accurate as possible, the time interval should be small enough to cover real-time fluctuations, but at the same time large enough so that it provides useful input to the user (e.g. an application) of the available bandwidth estimations.

It is also important to be aware of that the configuration of access capacities in commercial networks quite often allow traffic bursts in excess the specified capacity C. Thus, in small time scales the actual bitrate on the link level will be higher than C. If this in not taken into consideration when choosing T_p the resulting B_{Tp} time series will contain occurrences of negative values for available bandwidth.

Our approach to these challenges is a method based on performing stratified probing of the cross-traffic according to its detected profile. The rationale behind a strata oriented probing approach is to maximize the value of each probe packet sent by probing more during bursty cross-traffic periods than during almost idle periods.

The current version of our method is only applicable if the cross-traffic profile has a periodic component of significance. This type of cross-traffic is quite common due to the growing amount of services with video components on the Internet. The periodic and burst oriented nature of such services is described in more detail in Section II. In the case where cross-traffic does not have a periodic component of significance, and alternative probing approach should be considered.

B. Research Approach

In order to study both the feasibility of estimating available bandwidth through the use of stratified probing and also how well it performs, we chose an empirical research approach. To support this we established a hybrid active/passive measurement testbed which will be further described in Section V. The passive measurements are used to capture the real traffic profiles and available bandwidth, while the active measurements reflect the results when using our experimental implementation of the investigated method. The main service component used in our research as cross-traffic is adaptive video streaming which generates a typical periodical and burst oriented traffic profile. More details regarding adaptive video streaming will be provided in Section III.

C. Paper Outline

The structure of this paper is as follows. Section II provides an overview of related work; Section III provides characteristics of adaptive video streaming; Section IV describes our method of performing available bandwidth estimation; Section V describes our measurement setup; Section VI presents the results; Section VII presents a brief discussion; Section VII presents our conclusions and an outline of future work is given in Section IX.

II. Related Work

There are several approaches for estimating bandwidth along a network path, most of which fall into either the Probe Rate Model (PRM) or Probe Gap Model (PGM) categories [2]. The PRM approach is based on the principle of self-induced congestion and by this detecting available capacity, while the PGM approach uses observed inter-arrival time variations for probe packets to estimate the current level of cross traffic, which then combined with knowledge about the total capacity, can be used to produce an estimate of the currently available capacity. The stratified probing approach described and analyzed in our work, belong to the PGM category.

Within the range of PGM methods published over the last ten years, there are quite a few different versions [2]-[6]. They all use the principle of inserting probe packets in such a way that they follow the same path as the cross traffic of interest. However, when it comes to how many probe packets per time unit and patterns of such there are many differences.

In terms of how well the existing methods perform both in terms of accuracy and real-time capabilities a few studies have been published [7]. Early work in this area showed that a Poisson approach of spacing probe packets gave a significant improvement over the fixed approaches. More recent work [8] has documented the need for careful calibration of methods used in order for them to perform as good as possible. For the special case of available bandwidth estimation on access links, where the cross-traffic contains burst components – we have not been able to find any relevant research published. .

With regards to understanding the nature of adaptive video streaming and resulting traffic, we studied this from different perspectives in our earlier work [9] and [10]. Further on, in [11] we presented a new method for achieving a traffic shaping effect for adaptive video streaming, without involvement from network components. The purpose of this was to make the traffic easier to characterize by probing methods, by reducing the degree of burstiness in traffic.

In the first phase of the method we are presenting, we use a technique for detecting period and burst duration we earlier [12] proposed. This method uses serial correlation [13] on time series of observations in order to detect periodic properties in traffic. Such properties are commonly found in TCP traffic in general and in video streaming services in particular.

III. Adaptive Video Streaming

As stated in the introduction part there are many approaches to adaptive video streaming, and a new standard [1] has emerged in this domain. The specific solution used as basis for our research is the Smooth Streaming Solution from Microsoft [14]. The periodic nature of adaptive video streaming is given by the repeated requests for the next segment in a video stream, at a specific quality level (cf. Fig. 1).

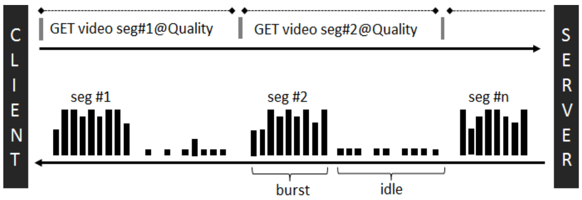

Figure 1. Traffic profile for adaptive video streaming

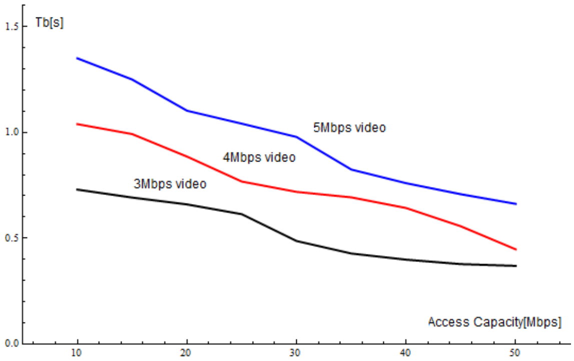

Figure 2. Passive measured burst period duration

The frequency of these requests differs between vendor solutions in general, and it is also to some extent dependent of implementation choices. The Smooth Streaming solution used in our experiments has a default GET segment request interval of 2 sec, which then would represent the period of interest (T_p) to identify for our probing method. The duration of the burst periods (time to get next video segment) varies depending on current quality level of the video stream, the access capacity and also the GET segment request interval. The resulting dynamics in duration of the burst period (T_b) is illustrated in Fig. 2. The presented values are based on passive measurements of a single video stream, operating without any other traffic present on the access link.

The presence of bursty traffic as described represents a challenge for both active and passive measurements. Passive measurements are of course easier as they do not have the need for any type of probing. However, even for passive measurements the resulting view of traffic load on an access link looks very different depending on over which period the average is calculated. In Fig. 3 we show the average bitrates for an adaptive video streaming service running at 4Mbps quality level, and the default T_p value of 2 sec.

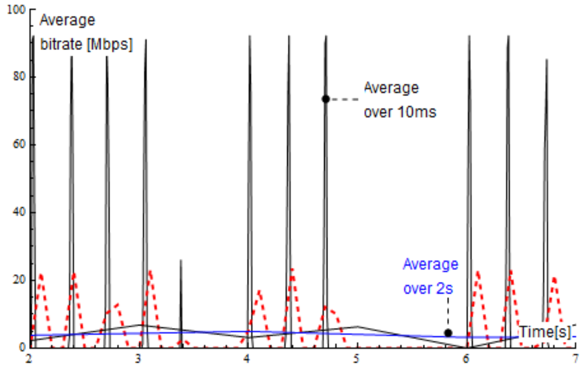

Figure 3. Passive measured average bitrates over different periods

The different lines in Fig. 3 give the results when using time intervals for average bitrate measurements in the range from 10ms to 2sec. The smallest time interval is able to capture the short lived traffic spikes up to about 90Mbps, while the largest time interval gives a more accurate view on the actual service quality level of about 4Mbps.

Comparing the passive measured bit rates in Fig. 3 with the traffic profile for adaptive video streaming in Fig. 1, the similarities are clear. During the burst periods there are a number of traffic spikes, while during the idle periods these spikes are not present. The size and number of spikes are closely related to streaming server capacity, physical link rate and basic TCP mechanisms. The specific TCP version used will also have an impact on the traffic pattern inside a burst period.

IV. Method

The suggested method for estimating available bandwidth is composed by different phases. How an actual implementation of the method will pass through these phases depends on the actual application. In a continuous estimation process of available bandwidth and a dynamic cross-traffic picture, there is a need to re-visit the first phase after some time. In the method flowchart (cf. Fig. 4) this is indicated by the return to first phase after *n* periods.

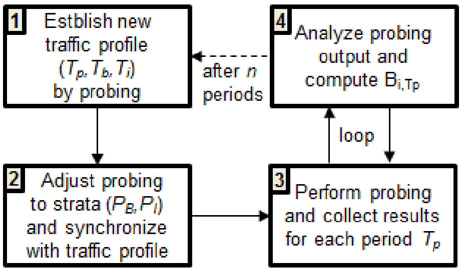

Figure 4. Phased approach for estimating available bandwidth

In the first phase, a traffic profile for the cross-traffic is established by use of active probing. This profile is used as input to the next phase where new probe rates are decided for each of the two strata (burst and idle sub-periods), and a synchronization of the probe strata and cross-traffic sub-periods are done. In the third phase the active probing of cross-traffic is performed in order to obtain v_j samples (cf. Eq. 3). In the last phase the available bandwidth is estimated.

A. *Establish Traffic Profile*

The periodic behavior in cross-traffic is described by the time parameter T_p and the duration of burst/idle periods are described by the time parameters T_b and T_i as shown in Fig. 5.

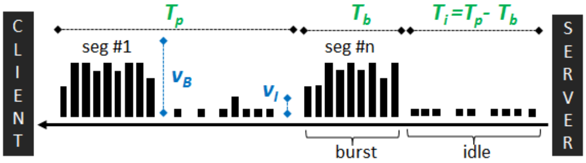

Figure 5. Burst and idle periods in cross-traffic

In our earlier work [12] we described and analyzed a method for estimating the relevant cross-traffic parameters by using active packet pair probing. This method is based on that the probe packet receiver observes the inter-arrival time IAT) between probe packet pairs, compares it with the cumulative average IAT and filter the samples based on this. If the sample is above the cumulative average, it is kept – if it is below, it is set to the average. The resulting time series of IAT observations ($t_{out,i}$) is used as input to computation of serial correlation up to a certain lag s (cf. Eq. 4). In our case where the packet pair probing period is known, the lag value maps directly over to the time dimension.

$$X_s = \frac{\sum_{i=1}^{N-s}(t_{out,i} - \overline{t_{out}})(t_{out,i+s} - \overline{t_{out}})}{\sum_{i=1}^{N}(t_{out,i} - \overline{t_{out}})^2} \qquad (4)$$

In order to understand what the output of the serial correlation X_s tells us, a graphical view is recommended. By using this, the parameters of interest (T_p, T_b) are possible to identify (cf. Fig. 6).

There are several things one could read out from a serial correlation performed over a time series with a periodical component. The example X_s output shown in Fig. 6 is based on a theoretical signal with a period of 2s, and where the signal inside this period has an idle part of 1.3s and a burst part of 0.7s. The burst part is not a square pulse.

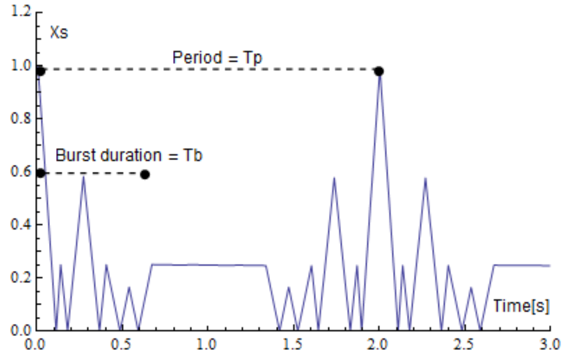

Figure 6. Serial correlation output for theoretical signal

The presence of peak values are indicators of periodic components, and visible side lobes around a peak is an indicator of that the periodic component is not a perfect square pulse, but rather composed of several pieces. By looking at the lower range in time for the X_s output, the width (or duration) of the burst part for each period is visible.

The amount of probe traffic required in this phase is very low. In our earlier work we showed that accurate estimates for both T_p and T_b were possible to obtain using probe packet rates (f_p) down to 160pps. With a probe packet size of 100bytes it corresponded to a bitrate of 128Kbps.

B. *Adjust Probing to Strata and Synchronize*

The direct output from the previous phase in terms of estimators for T_p and T_b are used in this phase in order to adjust the probing traffic according to the burst and idle periods. However, before the probe traffic is changed it is required to obtain some timing information which can be used to synchronize probing strata with the burst/idle periods in the cross-traffic.

The method we used for this purpose requires the presence of sequence number attached to each probe packet, and also the ability to restart the sequence numbering after a specific period.

Both these capabilities were supported by the probe traffic generator we used [15]. With this in place, the probe traffic used in the previous phase was changed so that it restarted its sequence numbers after the estimated T_p units of time, but keeping the same amount of probe packets per time unit (f_p). The probe packet sequence numbers were then available for use as an index (j), making it possible to summarize IAT observations ($t_{out,i,j}$) occurring at a specific time within each period (cf. Eq. 5) over *n* periods.

$$T_{out,n,j} = \sum_{i=1}^{n} t_{out,i,j} \qquad (5)$$

Further on, when performing this summarization for the whole index range *N*, it gives a the list $T_{Out,n}$ (cf. Eq. 6).

$$T_{Out,n} = \{T_{out,n,j,} 1 \le j \le N)\}$$
$$N = f_p * T_p \qquad (6)$$

The purpose of producing the $T_{Out,n}$ list is to see where within the probe sequence of length N the burst period starts. The required number of periods n for which the summarization is required performed in order to give this type of information is not obvious. One might think that a high n value is good, but as the results will show in Section X this is not entirely correct.

Having identified the time within a probe sequence where the burst period starts, we have also established a timing reference between our probing and the bounds for burst and idle periods in the cross-traffic. These bounds can then be used to implement the stratified probing, where we change the active probing from the continuous f_p rate over to $f_{p,b}$ during the cross-traffic burst period and $f_{p,i}$ during the idle period.

C. Active Probing and Result Collection

Choosing the optimal $f_{p,b}$ and $f_{p,i}$ values was outside of the scope for the research documented in this paper, and was left for future work. Thus, for the purpose of our experiments we chose a reference probe packet rate $f_{p,b}$=309pps for the burst period and set the $f_{p,i}$ to zero. The latter would not be appropriate in a scenario with a more complex cross-traffic mix than in our case, but sufficient to support the focus of this paper. The reference probe packet rate was used in the scenarios where the access capacity was at 10Mbps, independent of which quality level the adaptive video streaming (i.e. the cross-traffic service) was operating at. For the other access capacity levels, the $f_{p,b}$ was scaled up according to the changes in detected T_b. In other words, as T_b decreases (as a result of increased access capacity), $f_{p,b}$ was increased inversely proportional to this. This approach of scaling $f_{p,b}$ gave a constant number of probes during the burst period across all access capacity levels.

The probing pattern used in this phase differs slightly from the one used for establishing the cross-traffic profile. While in that phase we used packet pairs [12] with a certain gap between each pair, we use a continuous packet train (cf. Fig. 7) in this phase and collect IAT observations continuously. The reason for this change was to make better use of the information available in a probe packet sequence.

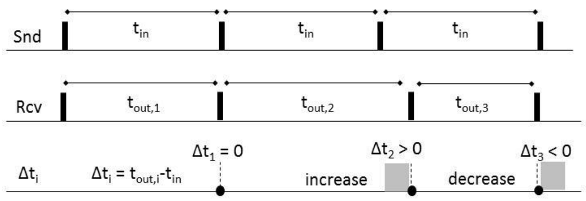

Figure 7. Probe packet IAT observations

As illustrated in Fig. 7, the Δt_i values represent the difference in spacing between probe packets as sent and received. Depending how the cross-traffic impacts the probe packets Δt_i can be positive or negative. An increase in spacing between probe packets is a certain indicator of cross-traffic impact, but even a reduced spacing observation may contain cross-traffic impact information. The required processing in order to extract all information available in the IAT observations ($t_{out,i}$) will follow in the next section.

D. Calculate Estimator for Available Bandwidth

For each $t_{out,i}$ observation, a cross-traffic delay component d_i is calculated and also a time shift element s_i. The latter is required in order for the subsequent calculation to be correct as a $d_i > 0$ would mean that the starting point for $t_{out,i+1}$ must be shifted. The continuous calculations are summarized as followed.

$$\begin{aligned} &if\ \Delta t_i = 0 \,\&\, s_{i-1} = 0 \text{ then } d_i = 0, s_i = d_i \\ &if\ \Delta t_i = 0 \,\&\, s_{i-1} > 0 \text{ then } d_i = s_{i-1}, s_i = d_i \\ &if\ \Delta t_i > 0 \qquad \text{ then } d_i = \Delta t_i, s_i = d_i \\ &if\ \Delta t_i < 0 \,\&\, s_{i-1} > 0 \text{ then } d_i = [\Delta t_i + s_{i-1}]^+ \\ &\qquad\qquad S_i = [d_i - s_{i-1}]^+ \end{aligned} \qquad (7)$$

Based on each d_i sample, a v_j sample can be calculated (cf. Eq. 8). This is used as input to the calculation of the total cross-traffic V_i (cf. Eq. 3) for a specific period. As per Fig. 5, the cross-traffic for the burst period is denoted V_B and for the idle period V_I. Since we are not probing during the idle period it gives that V_i=V_B.

$$v_j = C * d_i \qquad (8)$$

The estimator for available bandwidth $B_{i,Tp}$ during period i can then be calculated according to Eq. 3. In the results section we will also present the estimated burst rate $R_{i,burst}$ (cf. Eq. 9).

$$R_{i,burst} = \frac{1}{T_b} V_i[t - T_p, t] \qquad (9)$$

The reason why we also compute the burst rate is that we are interested in seeing how well the scaling of probe packet rate to $f_{p,b}$ based on burst period duration is performing. Alternative approaches for scaling it could have been based on access link capacity.

V. Measurement Setup

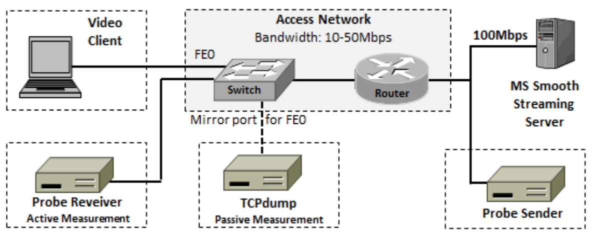

Figure 8. Hybrid active/passive measurement testbed

The purpose of the measurement testbed (cf. Fig. 8) is to provide both passive and active measurements of the cross-traffic generated across the access network. The passive measurements represents the real view and is collected by using TCPdump, while the active measurements represents the estimators obtained through stratified probing.

The testbed contains several components, ranging from the client side over to the server side. On the client side the video service is accessed by a dedicated Windows based PC, while the client receiving the probe traffic is on a separate Linux based PC. On the server side, we have the dedicated Microsoft Smooth Streaming server.

The access network part consists of a Cisco 2960 switch and a Cisco 1800 model router. Using this type of commercial off-the shelf components gives access to useful functions [16] for controlling bandwidth similar to those used in commercial networks.

The probe traffic sender and receiver are based on the CRUDE/RUDE tool [15], which provides the sufficient capabilities for generating traffic pattern based on trace files.

When running the experiments under different conditions (access capacity, video stream level etc) we use a measurement period of 10 minutes in order to collect both passive and active measurement results. Including some automated post processing, one measurement round covering all the different access capacities (10-50Mbps, with increments of 5Mbps) for a specific cross-traffic profile would then typically last for about 2 hours.

VI. Results

The results to be presented focus on demonstrating the capabilities of the method to support its four phases (cf. Fig. 4). The in depth analysis of each phase is covered in [12] and future work.

Typical results from each phase will be shown and explained. One specific scenario is being used through the following subsections and this is the case of a 5Mbps video stream as cross-traffic, configured with a segment request interval of 2sec and the access capacity set to 50Mbps.

A. Estimated Traffic Profile

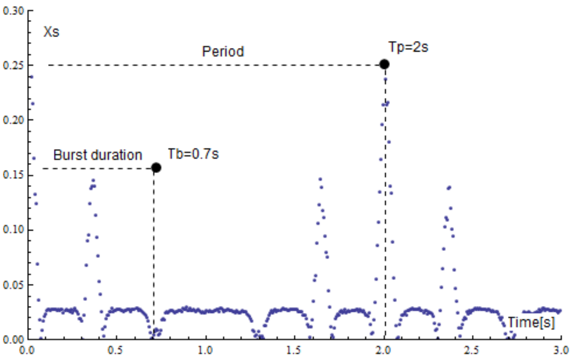

Figure 9. Period and burst

The active probing using in this phase was based on packet pairs. Each packet had a size of 100byte. The time between paired packets was 0.5ms and the time between packet pairs was 6.1ms. This gave a probe packet rate of about 300pps (240Kbps), which then produced 150 probe samples per second. The output from the serial correlation is shown in Fig. 9 where the lag parameter has been converted to time dimension. The adaptive video stream representing the cross-traffic in this case was operating at a 5Mbps quality level and a segment fetch interval of 2sec, across an access capacity of 50Mbps.

We can see how the peak value in the serial correlation output X_s is able to detect the period of 2sec in the cross traffic, which is correct. The observed side lobes are quite similar to the theoretical output as shown earlier in Fig. 6. The burst duration is also visible as the point where X_s goes to zero after the last side lobe in the lower end of the time scale at about 0.7s. This matches the passive measured burst duration as shown in Fig. 2.

As described in [12] it is easier to read out the burst duration from the serial correlation output when the duration is low (e.g. at 100Mbps access capacity). In the cases studied for access capacities between 10-50Mbps, a combination of serial correlation output and the method applied in the next phase of the method would be beneficial.

B. Synchronization of Probing Strata with Traffic Profile

In this phase we used the same active probing as in the previous phase, but now re-configured so that the sequence numbering of the probes are restarted after 600 probe packets corresponding to the estimated period T_p=2s. The synchronization point we are looking for is the offset into the sequence of 600 probe packets where the burst period starts.

As described in the method section, the required number of rounds n required for the $T_{Out,n}$ time series to give useful output was not obvious. Therefore, we have shown the output of this calculation for different n values in Fig. 10.

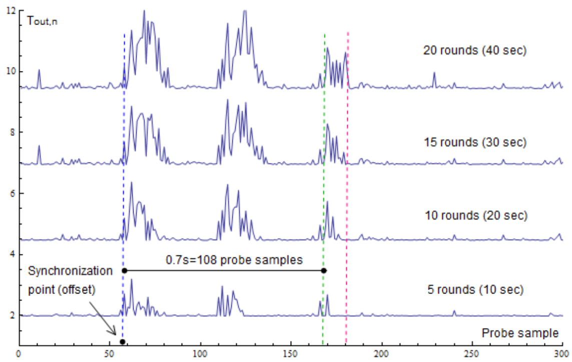

Figure 10. Synchronization of probe strata and burst period

As we see, the case when using the lowest n value of 5 which corresponds to sample size of 10sec is the one which gives the narrowest view of the burst duration. The time for the burst start is the same for all n values, but as n increases the period of interest is stretched. Thus, for the purpose of synchronizing our probing into different strata (burst and idle), all values of n gives the same starting point for the burst strata. However, for the purpose of providing an additional view on the burst duration – in order to simplify the interpretation of the serial correlation output, the lower n values are better.

C. Probe Traffic Scaled by Burst Duration

In this phase the probe traffic was reconfigured from the packet pairs used in the previous phases, to a sequence of probe packet equally spaced with parameter t_{in}. Since finding the optimal probing level was outside of the scope for this research we choose a starting point based on some basic experiments. The starting point chosen was for a 10Mbps access, for which we considered 2.5% of the bandwidth as acceptable to make available for active probing. This corresponds to the first entry in Table I, where a t_{in} value of 3.24ms is given for all cross-traffic cases.

TABLE I. BURST PERIOD PROBE TRAFFIC

Access [Mbps]	Probe Packet Size [byte]	Probe Packet Spacing, t_{in} [ms]		
		3M video	4M video	5M video
10	100	3.24	3.24	3.24
15	100	3.06	3.08	3.00
20	100	2.92	2.86	2.62
25	100	2.79	2.43	2.50
30	100	2.20	2.24	2.36
35	100	1.89	2.11	1.97
40	100	1.80	1.99	1.81
45	100	1.71	1.74	1.71
50	100	1.66	1.37	1.59

The scaling of the probe traffic according to burst period duration was done manually in our experiment as per the passive measured values in Fig. 2. The 3Mbps video probing was scaled based on the T_b profiles for this specific quality level, and similar for both 4Mbps and 5Mbps video.

D. Estimated Available Bandwidth

The available bandwidth estimations for the different scenarios were done based on time intervals of 600 seconds with active probing of cross-traffic. The range of scenarios covered was all combinations 3/4/5Mbps video streaming as cross-traffic, across all access capacity levels from 10-50Mbps.

In order to assess the accuracy of the results obtained through the active measurements (probing), we first present the passive measured (by TCPdump) burst bitrates $R_{i,burst}$ (cf. Eq. 9) for the 5Mbps video stream across a 50Mbps access (cf. Fig. 11). The majority of the measurement samples are located around 14Mbps, but there is also a portion located around 25Mbps. The sample mean for the whole 600s period is at 16.2Mbps.

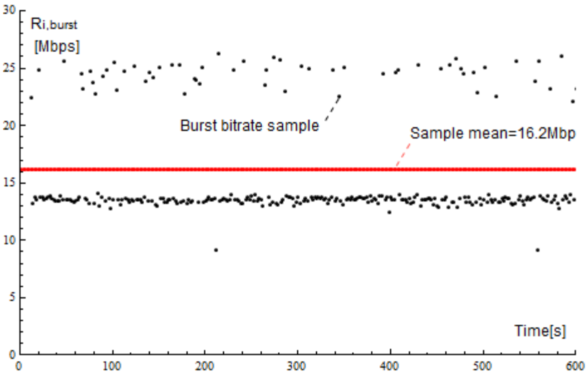

Figure 11. Passive measured burst bitrates

Moving over to active measurements by means of our probing during the burst periods of the cross-traffic we got the results as illustrated in Fig. 12. The spread in the measurement samples here are higher than in the case of passive measurements, and we also notice that the sample mean is at 17.9Mbps. The included plot of a 30sec moving average for the measurement samples shows a significant reduction in distribution spread. This gives us an indication on how fast our method is able to come up with a reasonable accurate estimation of burst bitrate, and thus also available bandwidth.

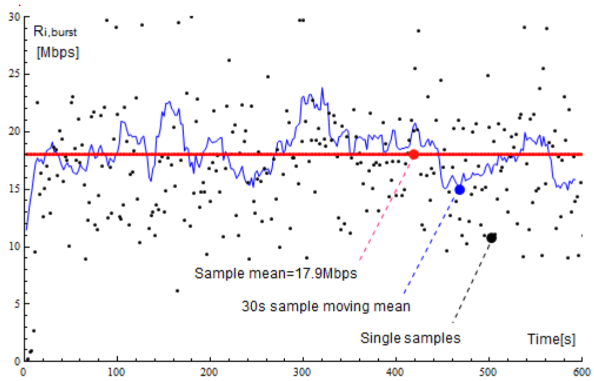

Figure 12. Active measured burst bitrates

In the following we take a look at how well our method performs over a wider range of access capacities, but still for the same 5Mbps video stream (cf. Fig. 13).

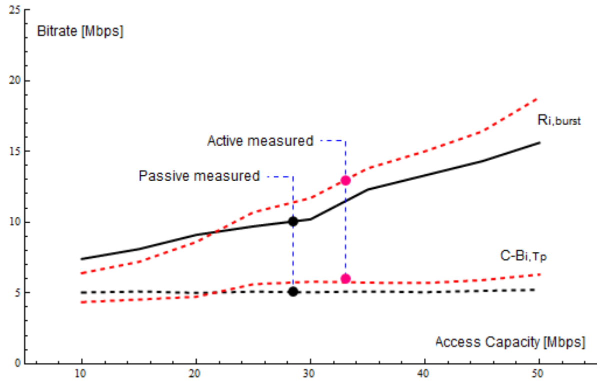

Figure 13. Comparison of passive and active measurement

We see here that the specific case we have presented in detail (5Mbps video on 50Mbps access) is actually the worst result for all capacity levels. For all other access capacity levels the difference between passive and active measured $R_{i,burst}$ and subsequently $B_{i,Tp}$ is smaller. For the purpose of making the illustration better the plot shows C-$B_{i,Tp}$ (estimated cross-traffic) rather than $B_{i,Tp}$ (estimated available bandwidth).

The differences between active and passive measurements are at worst in the order of 20% when looking at sample means over the 600s period. However, this should not be considered as a real measure of the accuracy of the method, but rather just a starting point. With more effort put into finding optimal probing patterns and rates, the accuracy is likely to improve further.

VII. Discussions

In our work we have made no attempt to compare the accuracy of our method against others. The main reason for this is that the published results for other methods and tools have not been addressing access links with burst traffic. Further on, those tools where source code are public available it would not be fair to test them for the sake of accuracy comparison in our scenario, as they were not made for this. It is also worth mentioning that most of these tools are about ten years old. However, it is worth mentioning that our stratified probing approach enables us to maintain a constant low probing rate even if the degree of burst duration is decreasing. This is the direct result of targeting the probes according to strata. In the scenario with 3Mbps video a cross-traffic the $f_{p,b}$ increases from the starting point of 309pps (247Kbps) up to 602pps (482Kbps), as the burst duration decreases from 0.72s to 0.37s. Looking at the f_p for the whole period T_p, remembering that $f_{p,i}$ is zero it is kept constant at 111pps (89Kbps) across all access capacity levels. The similar f_p values for respectively 4Mbps and 5Mbps video as cross-traffic are 160pps (128Kbps) and 210pps (168Kbps). We believe this clearly demonstrate the benefits of the stratified approach as we get more value for each probe sent.

In our work we have not used any specialized hardware to either generate traffic or to analyze it. All of the software used operates in user space and not kernel space of the operating systems. This directly implies that there is room for some errors in the results due to components performing multitasking during our experiments. We have tried to minimize the chances of this by following the advices found in [17] and also using dedicated nodes for each function in the experiments (cf. Fig. 8).

Our experimental approach to study our suggested method of stratified probing did not aim at developing a self-contained solution, which could be used outside a lab environment. However, there is a potential of doing so but it would require a certain amount of additional coding in the appropriate languages. This is outlined as part of future work, but it should be kept in mind that there are remaining technical challenges to be solved before it is recommended to invest this time. The main challenge in our view is to find a way to automatically choose a good starting point for the probe rates. In our work we used a level based on what we thought would be acceptable, but this assessment is of course highly subjective.

Even though we refer to our work as a new method for estimating available bandwidth on access links, we acknowledge that there are a lot of scenarios which are not covered by our approach. An example of this would be cases where periodic components in cross-traffic are not present at all. In such cases, our method does not add any value. In light of this, our method could be considered as a candidate component to be included in other more general methods.

VIII. Conclusions

The results from our empirical evaluation of the suggested method of applying a stratified approach for probing of cross-traffic are quite promising. We have showed that the different phases of our method are possible to implement when using a specific service type as cross-traffic. The choice of adaptive video streaming as the cross-traffic makes the findings quite relevant, reason being the growing amount of services with video components on the Internet.

The approach of using serial correlation as for analyzing time series of observations, such as IAT observations between probe packets is quite powerful. In this area we only provided a brief introduction to the concept, as we have presented this part in more detail earlier [12].

The benefit of stratifying the probe traffic according to the cross-traffic profile is quite clear. We have showed that our method is able to maintain about the same level of accuracy in the available bandwidth estimates over a wide range of burst degrees.

In order for our method to apply, there must be periodic behavior in the cross-traffic. This will not always be the case, but as the popularity of video services is growing we believe it will be quite common to see such behavior, especially on access links serving residential users.

A more complex traffic mix may of course change some of our results, but we believe that future methods in this domain should attempt to use the presence of periodic cross-traffic to their benefit in terms of improving accuracy in available bandwidth estimations.

IX. Future Work

The use of a more complex cross-traffic is interesting to study, in order to see how well our approach would perform in such a scenario. From one perspective it may add more complexity to the different phases of our method, but it may also contribute to smooth out sub-bursts and thereby simplify detection of burst durations. A burst period which is closer to a square pulse profile gives a clearer output when subject to serial correlation.

In order to further justify the gain of using knowledge about periodic traffic components as part of available bandwidth estimations, it would also be beneficial to collect and analyze traffic on a more aggregated level from real access networks.

Finding optimal probing patterns and rates was outside of the scope for our work. In order to take the validation of the method potentially one step further, we believe this would be an important area to investigate. We are especially interested in using single packet delay observations as basis for the active probing. Reason being that it could simplify the processing of observations on the receiver side, and also make the probing less vulnerable for packet loss and out-of-sequence events.

REFERENCES

[1] I. Sodagar, "The mpeg-dash standard for multimedia streaming over the internet," *Multimedia, IEEE*, vol. 18, no. 4, pp. 62 –67, April 2011.

[2] R. Prasad, C. Dovrolis, M. Murray, and K. Claffy, "Bandwidth estimation: Metrics, measurement techniques, and tools," *Network, IEEE*, vol. 17, no. 6, pp. 27 – 35, Nov.-Dec. 2003.

[3] J. Strauss, D. Katabi, and F. Kaashoek, "A measurement study of available bandwidth estimation tools," in *Proc. 3rd ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '03. New York, NY, USA: ACM, 2003, pp. 39–44.

[4] N. Hu, S. Member, P. Steenkiste, and S. Member, "Evaluation and characterization of available bandwidth probing techniques," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 6, pp. 879–894, 2003.

[5] V. Ribeiro, M. Coates, R. Riedi, S. Sarvotham, B. Hendricks, and R. Baraniuk, "Multifractal cross-traffic estimation," in *Proc. ITC Specialist Seminar on IP Measurement, Modeling, and Management*, 2000, pp. 15–1.

[6] E. Goldoni and M. Schivi, "End-to-end available bandwidth estimation tools, an experimental comparison," in *Traffic Monitoring and Analysis, ser. Lecture Notes in Computer Science*, F. Ricciato, M. Mellia, and E. Biersack, Eds. Springer Berlin Heidelberg, 2010, vol. 6003, pp. 171–182.

[7] G. Urvoy Keller, T. En Najjary, and A. Sorniotti. (January 2008). Operational comparison of available bandwidth estimation tools. *ACM SIGCOMM Computer Communications Review*. [Online]. 38(1). Available: http://www.eurecom.fr/publication/2413

[8] J. Sommers, P. Barford, and W. Willinger, "Laboratory-based calibration of available bandwidth estimation tools," *Microprocess. Microsyst.*, vol. 31, no. 4, pp. 222–235, 2007.

[9] B. J. Villa and P. E. Heegaard, "Improving perceived fairness and qoe for adaptive video streams," in *Proc. ICNS 2012*, March 2012, pp. 149-158.

[10] B. J. Villa, P. E. Heegaard, and A. Instefjord, "Improving fairness for adaptive http video streaming," in *Proc. EUNICE*, 2012, pp. 183–193.

[11] B. J. Villa and P. E. Heegaard, "Group based traffic shaping for adaptive http video streaming," in *Proc. 27th IEEE International Conference on Advanced Information Networking and Applications*, March 2013, pp. 830-837.

[12] B. J. Villa and P. E. Heegaard, "Detecting period and burst durations in video streaming by means of active probing," presented at the 3rd International Conference on Computer Communication and Management, Copenhagen, Denmark, May 2013.

[13] A. Rizk, Z. Bozakov, and M. Fidler, "H-probe: Estimating traffic correlations from sampling and active network probing," *CoRR*, vol. abs/1208.2870, 2012.

[14] A. Zambelli. IIS smooth streaming technical overview. Available: http://www.microsoft.com/silverlight/whitepapers/

[15] J. Laine, S. Saaristo, and R. Prior. Real-time udp data emitter (rude) and collector for rude (crude). Available: http://rude.sourceforge.net/

[16] C. H. Tim Szigeti, *End-to-End QoS Network Desi*. Cisco Press, November 2004.

[17] R. Prasad, M. Jain, and C. Dovrolis, "Effects of interrupt coalescence on network measurements," *Lecture Notes in Computer Science*, vol. 3015, pp. 247-256, 2004.

Bjørn J. Villa received his MSc degree in telematics at NTNU in 1995. After 15 years in the industry, he is now working on his PhD at NTNU. His research interest cover methods for optimizing QoS and QoE for Internet services, and he has published a number of research papers in this domain. A key service used in the research has been adaptive video streaming.

Poul E. Heegaard received his MSc degree in telematics at NTNU in 1989, and his PhD degree in 1998. Heegard is a Professor at ITEM/NTNU, and since 2009 he has been the head of the department. His research interests cover performance, dependability and survivability evaluation and management of communication systems.