Efficient Network-wide Probabilistic Available Bandwidth Estimation

Frederic Thouin
Elect. & Comp. Engineering
McGill University
Montreal, Canada
frederic.thouin@mail.mcgill.ca

Mark Coates
Elect. & Comp. Engineering
McGill University
Montreal, Canada
mark.coates@mcgill.ca

Michael Rabbat
Elect. & Comp. Engineering
McGill University
Montreal, Canada
michael.rabbat@mcgill.ca

ABSTRACT

In overlay and peer-to-peer applications it can be very useful to have accurate estimates of the maximum rate at which traffic can be sent along a network path without inducing significant congestion. The available bandwidth of the path provides an approximate measure of this rate. Existing approaches for generating an accurate estimate of available bandwidth require saturating the path for a short period of time with high-rate packet trains. When estimates are required for multiple paths, this strategy scales very poorly and induces an unacceptable measurement overhead. In this paper, we describe a distributed algorithm, based on Bayesian active learning and loopy belief propagation, for efficiently estimating the available bandwidths of multiple paths. We develop a probabilistic graphical model to capture the statistical dependencies between the available bandwidths of different paths; this allows us to learn about multiple paths each time we measure along a single path. Simulations and PlanetLab experiments indicate that this process requires few probes to generate accurate estimates.

1. INTRODUCTION

Many peer-to-peer and overlay applications could benefit from knowing the rate at which they could send an arbitrary amount of data along a path such that there is high probability that the output rate is (almost) the same as the ingress rate (or equivalently, such that the injected traffic induces minimal congestion). Peer selection in peer-to-peer streaming applications frequently involves some form of estimation of such a rate. Video streaming applications can use the information to choose transmission rates; the rate can also influence client-server association in content provider networks [6]. The available bandwidth of a path, which measures the unused capacity of the path over a specified time period, provides an approximate indicator of this rate (see [7] for a thorough discussion of the relationship).

The goal of this paper is to estimate the available bandwidths of multiple paths in a network. Our methodology exploits the correlations that arise when paths share links. Under certain modelling assumptions (see Section 2), the available bandwidth of a path is determined by a single *tight* link; each such tight link can decide the available bandwidths of multiple paths that traverse it. Measurements on one path thus can provide information not only about the available bandwidth of that path, but also the available bandwidths of other paths. We conduct measurement using packet-train probes and develop a probabilistic graphical model (factor graph [3]) to capture the dependencies between path (and link) available bandwidths. We adopt a sequential Bayesian learning methodology to infer marginal posteriors of the available bandwidths. With this approach, measurement noise can be treated in a more principled manner and active learning methods [1] can be employed to choose the paths to probe and the rates at which to measure, in order to minimize measurement overhead. The disadvantage is that we need to introduce measurement models (and priors), and there is the danger of modelling error. To mitigate this, we employ an empirical Bayes strategy, learning the parameters of a likelihood model from the measurements.

Related Work: In a short paper, it is impossible for us to do justice to the substantial literature on available bandwidth and its estimation. We refer the reader to [4] and [7] for much more complete discussions and surveys. The most effective techniques for available bandwidth estimation, including *Pathload* [5] and *pathChirp* [8], employ packet trains as the measurement methodology and are based on the principle of self-induced congestion. This principle implies that if a packet-train is sent at a rate exceeding the path's available bandwidth, then queuing will occur, and as a result, the egress rate is likely to be less than the ingress rate, and the inter-packet spacing is likely to increase over time. Conversely, if the rate of the input train is less than the available bandwidth, the egress rate will approximately equal the ingress rate and the inter-packet spacing will have no trend. The estimation approach is then to determine the rate at which the transition between these two behaviours occurs. The relationship between available bandwidth and this transition point is not exact; Liu et al. provide an excellent discussion on this point in [7].

Contributions: Existing bandwidth estimation techniques are designed for measurement of a single path. The methodology we present addresses estimation on multiple paths and is the first to employ probabilistic inference, graphical models, and active learning. In the broader context of network performance monitoring, Coates and Nowak introduced the use of graphical models for network tomography in [2], and

Rish et al. employed graphical models and active probing in [9], primarily for fault detection.

The rest of this paper is organized as follows. In Sect. 2, we formally define the available bandwidth metric. We describe our estimation methodology in Sect. 3. In Sect. 4, we present results obtained from simulations and online experiments on the PlanetLab network. In Sect. 5, we discuss other approaches we are exploring to solve this estimation task. Sect. 6 summarizes the contribution of the paper.

2. PROBLEM FORMULATION

Our goal is to estimate the available bandwidths of all paths in a network. The network is represented by a set of N directed links and a set of M paths. We assume that the routing topology of this network is known, and that it remains fixed for the duration of our experiments. Let r_l denote the ingress rate of a sequence of packets at link l, and let r'_l denote the egress rate. Also, let r_p denote the ingress rate and r'_p the egress rate of the end-to-end path p. We are interested in identifying the largest ingress rate at which we can transmit traffic along a path while maintaining an egress rate which is almost as large as the ingress rate with specified probability. More formally, for given $\epsilon > 0$ and $\delta > 0$, we seek the largest r_p such that $\Pr(r'_p > r_p - \epsilon) > 1 - \delta$. This probability is defined on the field of events consisting of all sequences of packets injected at rate r_p over some period within the measurement interval. We call this maximal rate r^*_p the (ϵ, δ)-available bandwidth for path p. We stress that this is different from the standard definition of available bandwidth, which is framed in terms of utilization and capacity, but it does more directly reflect our quantity of interest[1]. We want to know the rate at which we can inject traffic so that with some specified probability the egress rate will almost equal the ingress rate.

We wish to establish a relationship between the (ϵ, δ)-available bandwidth for a path and the (ϵ, δ)-available bandwidths of its constituent links. Consider an individual link. Given an ingress rate r_l, suppose there exists a small constant $\epsilon_l > 0$ and a small constant $0 < \delta_l < 1$ such that $\Pr(r'_l \leq r_l - \epsilon_l) \leq \delta_l$; that is, the probability that the egress rate deviates significantly from the ingress rate is bounded by δ_l. We can determine a similar relationship for an end-to-end path with link set L_p via the union bound, $\Pr(\cup_{l \in L_p} \{r'_l \leq r_l - \epsilon_l\}) \leq \sum_{l \in L_p} \delta_l$. From this relationship, it follows that

$$\Pr(\cap_{l \in L_p} \{r'_l > r_l - \epsilon_l\}) = \Pr(r'_p > r_p - \sum_{l \in L_p} \epsilon_l) > 1 - \sum_{l \in L_p} \delta_l.$$

The (ϵ, δ)-available bandwidth for path p is then the maximum r_p subject to $\Pr(r'_p > r_p - \sum_{l \in L_p} \epsilon_l) > 1 - \sum_{l \in L_p} \delta_l$, where $\sum_{l \in L_p} \epsilon_l < \epsilon$ and $\sum_{l \in L_p} \delta_l < \delta$. We denote the largest such ingress rate by $r^*_p(\epsilon, \delta)$. Similarly, at the link level, for given ϵ_l and δ_l, let $r^*_l(\epsilon_l, \delta_l)$ denote the largest r_l such that $\Pr(r'_l > r_l - \epsilon_l) > 1 - \delta_l$.

Our inference procedure builds on the assumption that on each path there is a tight link $l^* \in L_p$ which essentially determines the (ϵ, δ)-available bandwidth on the entire path. That is, supposing that r^*_p is the (ϵ, δ)-available bandwidth on path p, we assume there is a link l^* such that $\epsilon_{l^*} = \epsilon \gg \epsilon_l$ and $\delta_{l^*} = \delta \gg \delta_l$ for all $l \in L_p$, $l \neq l^*$ at input rate r^*_p. Consequently, $\sum_{l \in L_p} \epsilon_l \approx \epsilon_{l^*} = \epsilon$ and $\sum_{l \in L_p} \delta_l \approx \delta_{l^*} = \delta$, and thus $r^*_p(\epsilon, \delta) \approx r^*_{l^*}(\epsilon, \delta)$. From this perspective, we can see that $r^*_p(\epsilon, \delta) \approx \min_{l \in L_p} r^*_l(\epsilon, \delta)$. Another way to view our assumption is that the input rate r^*_p is well below the available bandwidth on each of the non-tight links $l \neq l^*$, so it is possible to select $\epsilon_l \approx 0$ and $\delta_l \approx 0$ while still ensuring $\Pr(r'_l > r_l - \epsilon_l) > 1 - \delta_l$ for $r_l \approx r^*_p$.

3. METHODOLOGY

Our approach to (ϵ, δ)-available bandwidth estimation is based on packet train measurements. Each packet train probes one path at a specified rate. When we probe path p at a rate r, the outcome of the measurement is a binary variable $z = \mathbf{1}\{r' > r - \epsilon\}$, indicating whether the output rate r' was within ϵ of the input rate.

We employ a probabilistic graphical model (factor graph [3]) to capture statistical dependencies. The factor graph contains one variable node, y_p, for each path and one variable node, x_l, for each logical link[2] in the network. Both link and path variables are modelled as discrete random variables with, e.g., $\Pr(y_p = r)$ being the probability that the (ϵ, δ)-available bandwidth on path p is r. The variable nodes are connected to function nodes expressing the relationship $y_p = \min_{l \in L_p} x_l$; thus, there is one factor node for each path variable. When a new measurement, z, is obtained, we update the factor graph by running belief propagation [11]. We adopt the likelihood model $L(z = 1 | y_p, r) = \text{logsig}(-\alpha(r - y_p))$ for the measurements[3], where α is a small positive constant. We have chosen this model based on experimental data; for each network, the parameter α is estimated via standard regression techniques by performing multiple measurements on the constituent paths at various rates r. In our experience, the likelihood model is relatively stable so the training exercise need only be executed rarely. On the other hand, every time the available bandwidth estimation algorithm is executed, new data are collected and the α value can be easily updated.

Instead of attempting to evaluate the full posterior distribution of all path available bandwidths after each measurement, we choose to propagate marginal posterior distributions for the link available bandwidths (the path available bandwidth is a deterministic function of the available bandwidths of its constituent links). We adopt this approximation because the true posterior cannot be expressed analytically and occupies a high-dimensional space. Attempting

[1]Most existing available bandwidth estimation techniques effectively identify the quantity we have defined (for $\epsilon = 0$ and $\delta = 0$). Fluid models of traffic have been employed to argue for the equivalence of this rate-based quantity and the utilization-based available bandwidth metric [5]. Liu et al. provide a thorough analysis and some experimental results that demonstrate that the equivalence is only approximate [7]. We choose not to introduce a new name for this probabilistic, rate-based quantity because there is an approximate equivalence and because it has been the focus of most "available bandwidth" estimation techniques.

[2]Recall that we assume the routing topology is known. Rather than working with the routing topology directly, we reduce it to the minimal equivalent logical topology by combining links that are in series. Under our modelling assumptions, the available bandwidth of two or more links in series is equal to the minimum available bandwidth of the constituent links. Thus, operating on the logical topology loses no information.

[3]To be exact, we bound this function to lie in the range $[\kappa, 1 - \kappa]$ for a small constant κ to ensure that we assign some likelihood to unexpected measurement outcomes at all ingress rates.

to track the full posterior with any accuracy thus involves a very high computational expense.

We would like to obtain network-wide available bandwidth estimates using as few probes as possible. To accomplish this, we adopt an active learning approach, sequentially choosing which path to probe next based on the measurements already obtained. Given that we will probe a path p next, we choose the probing rate to be the median of the marginal $P(y_p = r|\mathbf{z})$, since this is the rate of highest uncertainty; i.e., the rate at which we are equally likely to observe a 1 or a 0 according to our current model. Probing at the median in this fashion is equivalent to conducting a probabilistic binary search for the available bandwidth on path p [1]. The other major design issue to be addressed is how to choose which path to probe next. A naïve approach would be to probe each path in round robin fashion. However, this would result in often probing paths for which we already have relatively confident estimates. Instead, we consider two data-driven, randomized approaches to path selection. The weighted entropy approach selects the next path to probe with probability proportional to the entropy of its current marginal distribution. The weighted confidence interval approach selects the next path to probe with probability proportional to the width of the bandwidth range (in Mbps) required to encapsulate η percent of the probability mass. Both schemes favour probing paths for which the current available bandwidth estimate has higher uncertainty.

We specify the stopping criterion in terms of confidence intervals — we require that η percent of the probability mass lies in a bandwidth range smaller than β for all paths. In the deployed software, we also specify a maximum number of measurements per path, to address the case of high variability where convergence is difficult to achieve.

4. RESULTS

In this section we describe preliminary results of simulations and experiments on the PlanetLab network[4]. In all cases, we focus on the (ϵ, δ)-available bandwidths for $\epsilon = 5$ Mbps and $\delta = 0.5$. The purpose of the simulations is to explore the efficacy of our proposed learning strategies. These are not network simulations, so they do not test modelling assumptions at all (that is the purpose of the PlanetLab experiments). The simulations are conducted in Matlab, on a topology derived (using traceroute[5]) from PlanetLab. The simulation topology consists of $M = 20$ paths and $N = 32$ links. The link available bandwidths are assigned using a uniform distribution over the range $[1, 100]$ Mbps. Probe outcomes are generated according to our likelihood model.

Figure 1 compares the three path selection algorithms outlined in the previous section (Round-robin (RR), Weighted Entropy (WE), and Weighted Confidence Interval (WCI)). We show the number of measurements per path required for the algorithm to terminate (averaged over 25 runs with the same topology but different random link capacities), as well as the accuracy (an estimate is considered accurate if the real available bandwidth lies within the identified confidence interval). For comparative purposes, we also show the average number of measurements and accuracy required

[4]http://www.planet-lab.org/

[5]Topology estimation using traceroute can inflate the number of routers and identify non-existing links [10], but it is sufficiently accurate for preliminary experimentation.

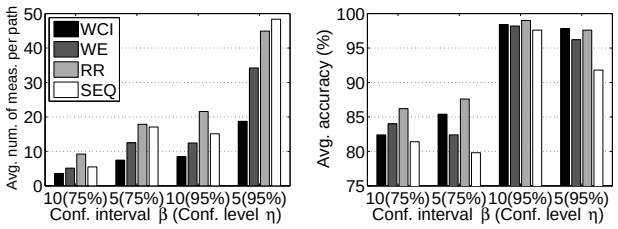

Figure 1: Simulation results: measurements required and accuracy achieved. Results are averaged over 25 runs for different confidence levels η and intervals β. The topology has 20 paths and 32 links.

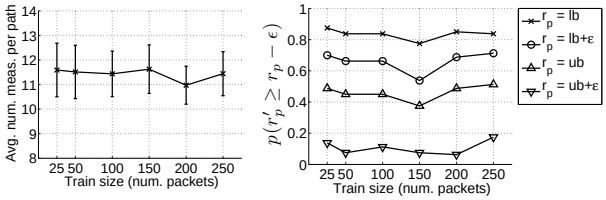

Figure 2: Accuracy of our estimates and number of measurements per path averaged over 20 runs of 16 tests as a function of the train size used. We used $\beta = 10$, $\eta = 0.95$, $\epsilon = 5$ and WCI for path selection.

when our active learning algorithm is run independently and sequentially on each path (SEQ). In most cases, SEQ requires fewer measurements than the round-robin strategy with the graphical model. This is due to the fact that not all paths require the same number of measurements; in the RR case, the algorithm iterates through all paths even the ones that have already met the required confidence criterion, which is not the case in SEQ. This shows that employing the factor graph in a naive manner can be disadvantageous because noisy measurement outcomes can spread uncertainty to other path estimates. Both data-driven approaches significantly reduce the number of measurements required.

For our online experiments, we deployed our measurement software on various nodes on the PlanetLab network. The online topology consists of $M = 30$ paths and $N = 65$ links; all possible end-to-end paths between six selected nodes. To construct the factor graph using the method described in Sect. 3, we first extract the topology (using traceroute). We do so before every experiment, but noticed that the routing table was relatively stable for a given set of nodes over the days during which we ran the experiments. In Fig. 2, we examine the estimation performance and explore the impact of using different train sizes. The left panel of the figure shows the number of measurements required to achieve the required confidence level (β=10Mbps, $\eta = 0.95$). The right panel shows the outcome of transmission tests conducted at the end of the estimation interval. We performed four tests on four disjoint paths (for a total of 16 tests per run) by sending trains of 2000 packets of 1000 bytes and observing the output rate. The trains were sent at four different rates — the lower bound of the confidence interval *lb*, the lower bound plus 5 Mbps (approximately the midpoint of the confidence interval), the upper bound of the confidence interval *ub*, and 5 Mbps above the upper bound. Based on these tests, we calculate the empirical probability that the egress rate exceeds the ingress rate less the tolerance factor ϵ. The

performance varies very little with the number of packets in the train, indicating that, for this network at least, 25 packets per train would suffice. Probing at the upper bound of the confidence interval results in an empirical probability close to 0.5, the target δ for our experiments. When the upper bound is exceeded by a few Mbps, the empirical probability drops to values around $0.1 - 0.2$, and probing at the lower bound leads to an empirical probability of $0.8 - 0.9$. These results indicate that the estimated confidence intervals provide a very good indication of the rate at which data can be transmitted with reasonable probability of avoiding congestion. The technique slightly underestimates the available bandwidths; this is probably due to an asymmetry we observe in the uncertainty ("noise") of measurement outcomes, depending on whether the input rate is less than or exceeds the available bandwidth. This is not reflected in our symmetric likelihood model.

5. DISCUSSION

The estimation approach we have outlined involves each probe measuring a single rate (with a decision about the rate made using an active learning strategy). An alternative measurement strategy, employed previously in available bandwidth estimation for a single path [8], is to test at multiple rates with each probe. This is achieved by varying the spacing between packets in the probe-train, so that the rate (number of bytes/time) changes for the first $k+1$ packets compared to the first k. In one measurement we can make multiple binary tests, constructing a value $z(k)$ for each value of $k \in K_{\min}, \ldots, K$. Here K is the total number of packets in the probe and $K_{\min}$ is the minimum number of packets required to generate a meaningful rate test.

The chirp approach has a clear advantage of providing information about multiple rates with one measurement, but for a fixed byte budget, the trade-off is that each binary test is much noisier. Moreover, constructing a likelihood function for the multiple outcomes of a chirp must be approached with care, because the "noise" is correlated. The successive tests in the chirp are affected by the same competing traffic, so the validity of the independence assumption between measurements is much more questionable. Given this issue, the question arises of how we can meaningfully incorporate the richer information provided by the chirp probes.

One method is to view the outcome of the chirp probe not as a series of binary tests, $z(k) = \mathbf{1}\{r'_p(k) > r_p(k) - \epsilon\}$, but as a single rate outcome $z = r(k^*)$. In the noiseless, ideal case, we would like to choose k^* such that $r'_p(k) = r_p(k)$ for all $k < k^*$ and $r'_p(k) < r_p(k)$ for all $k > k^*$. In the noisy setting, we only expect these conditions to hold for most k; we choose k^* to maximize how often the conditions hold.

We thus have specified a measurement z and can strive to identify an appropriate likelihood function $L(z|r^*_p, \mathbf{r_p})$. The likelihood is dependent on both the available bandwidth r^*_p and the probe structure, identified by the vector of probing rates $\mathbf{r_p}$. We are currently conducting experiments to identify a suitable parametric model for the likelihood, which can then be trained in a similar fashion to the likelihood function for single-rate packet trains.

In contrast to*pathChirp*, we propose to employ an adaptive measurement process, adjusting the range and spacings of the chirp probes based on similar active learning principles to those we have used for single-rate measurements. We anticipate that a hybrid measurement approach will prove most effective, where chirps are used initially to quickly locate a small range of probable values for the available bandwidth and then single-rate packet trains, which have less noise, are used to provide fine-grained resolution.

6. CONCLUSION

We have introduced a method for the estimation of the available bandwidths of multiple paths. We describe the statistical dependences between the available bandwidths of different paths using a probabilistic graphical model. The estimation strategy follows a Bayesian learning framework and involves the application of loopy belief propagation to infer marginal posteriors. We use active learning methods to choose the path to probe and the probing rate in order to minimize the measurement overhead. Simulations indicate that the active learning strategies significantly reduce the number of probes required to form an estimate. On-line experiments on the PlanetLab network show that the estimates our methodology generates provide a very good indication of the maximum rates at which data can be transmitted so that there is small probability of inducing congestion.

7. REFERENCES

[1] R. Castro and R. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, Jul. 2008.

[2] M. J. Coates and R. Nowak. Networks for networks: Internet analysis using graphical statistical models. In *Proc. IEEE Workshop on Neural Networks for Signal Processing*, Sydney, Australia, Dec. 2000.

[3] B. Frey. *Graphical models for machine learning and digital communication*. MIT Press, 1998.

[4] M. Jain. *End-to-end available bandwidth estimation and its applications*. PhD thesis, Georgia Institute of Technology, May 2007.

[5] M. Jain and C. Dovrolis. End-to-end available bandwidth: measurement methodology, dynamics, and relation with TCP throughput. *IEEE/ACM Transactions Networking*, 11(4):537–549, Aug. 2003.

[6] M. Jain and C. Dovrolis. Path selection using available bandwidth estimation in overlay-based video streaming. *Comp. Networks*, 52(12):2411–2418, Aug. 2008.

[7] X. Liu, K. Ravindan, and D. Loguinov. A stochastic foundation of available bandwidth estimation: Multi-hop analysis. *IEEE/ACM Trans. Networking*, 16(1):130–143, Feb. 2008.

[8] V. Ribeiro, R. Riedi, R. Baraniuk, J. Navratil, and L. Cottrell. pathchirp: Efficient available bandwidth estimation for network paths. In *Proc. Passive and Active Measurements Conf.*, La Jolla, CA, Apr. 2003.

[9] I. Rish, M. Brodie, S. Ma, N. Odintsova, A. Beygelzimer, G. Grabarnik, and K. Hernandez. Adaptive diagnosis in distributed systems. *IEEE Trans. Neural Networks*, 16(5):1088–1109, Sep. 2005.

[10] R. Sherwood, A. Bender, and N. Spring. Discarte: A disjunctive internet cartographer. In *Proc. SIGCOMM*, Seattle, WA, Aug. 2008.

[11] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *Proc. Advances Neural Information Processing Systems (NIPS)*, Denver, CO, Dec. 2000.