

# Tuning the Temporal Characteristics of a Kalman-Filter Method for End-to-End Bandwidth Estimation

Erik Hartikainen<sup>1,2</sup> and Svante Ekelin<sup>2,3</sup>

<sup>1</sup> Dept. of Science and Technology, Linköping University, Norrköping, Sweden

<sup>2</sup> Network Control Lab, Ericsson Research, Stockholm, Sweden

<sup>3</sup> Dept. of Computer Science and Electronics, Mälardalen University, Västerås, Sweden

**Abstract**—In this paper we present a way of tuning the temporal characteristics of a new available-bandwidth estimation method, BART. The estimation engine in this method is Kalman-filter based. A current estimate of the available bandwidth is maintained, and for each new sequence of probe packet pairs an updated estimate is produced. The main input parameters needed by the Kalman filter are the variance of the measurement noise and the covariance of the process noise. The former is measured by the method, whereas the latter is not in general attainable by analytical or empirical investigation. Instead, it is reasonable to treat this as a tunable parameter. We discuss how the temporal characteristics of the tracking of end-to-end available bandwidth may be tuned.

## I. INTRODUCTION

### A. Overview

The capability of estimating end-to-end available bandwidth is useful in several contexts, including service level agreement verification, network monitoring and server selection. Estimation of bandwidth in real-time, with per-sample update, opens up for many applications, including adaptation based on available bandwidth directly (rather than measures such as loss or delay) in e.g. congestion control and streaming of audio and video.

Estimation methods based on a filtering approach possess this compelling feature of producing an updated estimate for each new sampling of the system properties. For instance, by probing the network with probe packets once every second, one obtains an updated estimate of the currently available bandwidth every second, allowing to track the bandwidth on a second-by-second basis, provided of course one has been able to successfully apply a filter method to the estimation problem. Recently, such a method, BART (Bandwidth Available in Real-Time), has been evaluated [1], including measurements over the Internet.

In this paper, we investigate the possibility of tuning BART to track the available bandwidth at an arbitrary averaging time scale  $\tau$  [2]. There is no canonical time scale at which to define available bandwidth. The relevant time

resolution may be different for various applications.

The time scale of bandwidth tracking offered by BART is related to two adjustable properties. The obvious one is the inter-probing time. In the example above, when we probe once per second, we cannot hope to accurately track the bandwidth fluctuations at any time resolution better than approximately two seconds (cf. the sampling theorem in information theory). By decreasing the inter-probing time, an improved tracking performance is expected; however, the trade-off is the larger amount of probe traffic affecting the network. The more subtle adjustable property is the process noise covariance,  $Q$ , one of the crucial filter parameters. This is a  $2 \times 2$  matrix in the BART formalism.

In the present paper, we explore how the elements of  $Q$  could be chosen so as to optimize the tracking performance of BART for desired bandwidth variability. We introduce and apply a specific variability measure, which captures the effect of both time resolution and traffic aggregation.

### B. Related Work

Several other bandwidth estimation methods have been proposed [3, 4, 5, 6, 7, 8, 9, 10]. These methods have been compared for performance [11, 12].

In a seminal paper on packet-pair techniques [13], Keshav discussed using a Kalman filter for the estimation of “bottleneck service rate” for end-point flow control purposes, and concluded that this would not be practical. However, his analysis rested on the assumption that queuing service in network nodes is based on stateful flow-based round-robin instead of stateless first-come first-served (FCFS), whereas typically the opposite holds in today’s Internet.

## II. FILTER-BASED BANDWIDTH ESTIMATION

### A. Filter-Based Estimation

In a filter-based approach, the state of a system is estimated from repeated measurements of some quantity dependent on the system state, given models of how the system state evolves from one measurement occasion to the next, and how the measured quantity depends on the system state. Both these dependencies include a random noise term; the process noise and the measurement noise, respectively.

The system equations are then<sup>1</sup>

$$x_k = f(x_{k-1}) + w_{k-1} \quad (1)$$

$$z_k = h(x_k) + v_k \quad (2)$$

where  $x$  is the state of the system,  $z$  is the measurement,  $w$  is the process noise and  $v$  is the measurement noise. The functions  $f$  and  $h$  represent the system evolution model and the measurement model, respectively. The subscript refers to the “discrete time”.

A *filter* is a procedure which takes a previous estimate  $\hat{x}_{k-1}$  and a new measurement  $z_k$  as input, and calculates a new estimate  $\hat{x}_k$  of the system state. A compelling property of filters is that they are capable of producing estimates in real-time, i.e. tracking the system state. For each new measurement, the previous estimate is updated.

A key point of this paper is how the relative weights of the previous estimate and the new measurement in the filtering process influence the temporal characteristics of the method. We return to this below.

If the functions  $f$  and  $h$  are linear, and if both the process noise and the measurement noise are Gaussian and uncorrelated, there is an optimal filter, namely the Kalman filter [14]. Experience has shown that Kalman filters often work very well, even when these conditions are not strictly met. Another important advantage with Kalman filters is that, unless the dimensionality of the system is very large, they are computationally light-weight, with minimal requirements on CPU and memory.

In this linear case the system equations can be expressed using matrices:

$$x_k = Ax_{k-1} + w_{k-1} \quad (3)$$

$$z_k = Hx_k + v_k \quad (4)$$

The Kalman filter equations allowing calculation of the new estimate from the previous estimate and the new measurement are (for details, see [14]):

$$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-) \quad (5)$$

$$P_k = (I - K_k H)P_k^- \quad (6)$$

where

$$\hat{x}_k^- = A\hat{x}_{k-1} \quad (7)$$

$$P_k^- = AP_{k-1}A^T + Q \quad (8)$$

and

$$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1}. \quad (9)$$

Kalman filtering can be understood as a process where there are two phases of calculation in each iteration. First, there is a “prediction” phase, where the previous estimate evolves

one discrete time step according to the system model (7). Then, there is a “correction” phase, where the new measurement is taken into account (5). One also computes the updated error covariance matrix  $P_k$  of the state estimate.

Of special interest in these equations is the Kalman gain  $K_k$ , given by (9) and appearing in (5) and (6). This can be interpreted as the relative weight given to the new measurement as opposed to the pure expected evolution of the previous estimate.

As can be seen from (8) and (9), the Kalman gain increases with  $Q$  and decreases with  $R$ . These required inputs to the Kalman filter are the covariances of the process noise  $w$  and measurement noise  $v$ , respectively. An intuitive understanding of the importance of these quantities may be acquired by the following arguments:

- Large variations of the noise in the system model (high  $Q$ ) mean that the prediction according to the system model is likely to be less accurate, and the new measurement should be weighted heavier.
- Large variations in the measurement noise (high  $R$ ) mean that the new measurement is likely to be less accurate, and the prediction should be weighted heavier.

#### B. The BART Filter Method for Available Bandwidth Estimation

The BART (Bandwidth Available in Real-Time) method for available bandwidth estimation [1] makes use of a Kalman filter in order to produce an updated estimate of the available bandwidth over a network path for each new measurement. Measurements are performed at randomized probing rates, in order to achieve optimal statistical estimation properties. Each measurement consists of sending a sequence of pairs of packets, which are time-stamped on sending and on arrival, and calculating the average relative increase  $\varepsilon$  in packet-pair time separation. At the same time, the variance of  $\varepsilon$  is computed, so  $R$  is measured. From a simple FCFS network model [10], it is seen that the expectation value of  $\varepsilon$  is zero when the probing rate  $u$  is less than the available bandwidth  $B$ , and grows in proportion to the overload when the probing rate is larger:

$$\varepsilon = \begin{cases} 0 & (u < B) \\ \alpha(u - B) = \alpha u + \beta & (u \geq B) \end{cases} \quad (10)$$

In order to use dimensionless quantities in the BART filtering, we normalize  $u$  with respect to the maximum probe intensity  $u_{max}$ .

It should be noted that the Kalman filter method is very “forgiving”, and good results are often produced even when the ideal conditions are slightly broken. So, even if a real system displays characteristics which deviate somewhat from this piecewise linear system curve, the resulting available bandwidth estimate is not automatically invalidated.

Of course, all variables in this model are dynamical, i.e. may vary in time, so they depend on the subscript (which is sometimes suppressed in this exposition).

<sup>1</sup> This can be formulated more generally, when the system is also influenced by a control input. Since this is not needed in BART, this additional degree of freedom is not used in this paper.

This model allows for application of a Kalman filter, when we represent the state of the system by a vector containing the two parameters of the sloping straight line

$$x = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \quad (11)$$

We may then write, for the measurement  $z_k$  of the strain at discrete time  $k$ ,

$$z_k = Hx_k + v_k \quad (12)$$

where

$$H = [u \quad 1]. \quad (13)$$

Also, we write for the evolution of the system state

$$x_k = x_{k-1} + w_{k-1} \quad (14)$$

which means that we may apply the Kalman filter formalism with  $A = I$ .

The Kalman equations in BART become computationally very simple, and only a few floating-point operations are needed at each iteration.

When the filter estimates the system state variables  $\alpha$  and  $\beta$  we immediately obtain the estimate for the available bandwidth  $B$ , since these quantities are related by

$$\alpha B + \beta = 0. \quad (15)$$

### C. The Structure of $Q$

The state vector  $x$  in BART is two-dimensional, so the covariance  $Q$  of the process noise  $w$  is a  $2 \times 2$  matrix.

In (14), the process noise is defined as the change of the system state between two consecutive measurements (which is due to the change in cross traffic).

$$w = \Delta x = \begin{bmatrix} \Delta\alpha \\ \Delta\beta \end{bmatrix} \quad (16)$$

so we have

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} = \begin{bmatrix} V(\Delta\alpha) & C(\Delta\alpha, \Delta\beta) \\ C(\Delta\alpha, \Delta\beta) & V(\Delta\beta) \end{bmatrix}. \quad (17)$$

$Q$  is a symmetric matrix, i.e.  $Q_{21} = Q_{12}$ , so there are three degrees of freedom available for tuning. The allowed ranges are given by the inequalities  $Q_{11} \geq 0$ ,  $Q_{22} \geq 0$ , and  $-\sqrt{Q_{11}Q_{22}} \leq Q_{12} \leq \sqrt{Q_{11}Q_{22}}$ . These bounds are simply due to the properties of variances and covariances of stochastic variables.

It should be noted that in this model,  $\alpha$  is the inverse of the tight link capacity. When the cross traffic intensity varies, we can expect  $\beta$  to vary and  $\alpha$  to be rather stable. This means that  $Q_{11}$  and  $Q_{12}$  could be expected to be small compared to  $Q_{22}$ , unless the network is in a state where the tight link is frequently moving.

### D. Tuning BART

The BART estimation method comprises several tunable components:

- The probe packet size
- The number of probe-packet pairs in each measurement
- The organization of the probe pairs, i.e. separated or as a train
- The inter-measurement time separation
- The probability distribution for the probing intensity
- The covariance matrix of the process noise

In the present paper, we are specifically addressing the temporal characteristics of BART, and we will limit our attention to tuning of the latter component, i.e.  $Q$ , the covariance matrix of the process noise. All the other components are fixed in this study, although some of them have been evaluated in [1]. There is plenty of opportunity for further performance increase. However, the qualitative dependence of the temporal characteristics on  $Q$  we find in the present paper is likely to remain also when more evolved choices of other components are made.

## III. EXPERIMENT

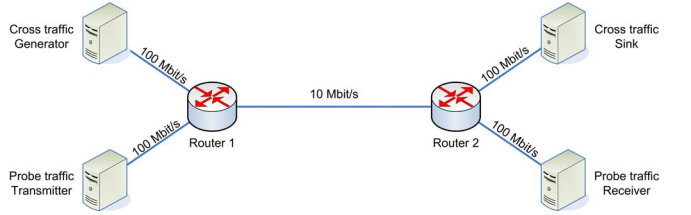


Fig. 1. An illustration of the testbed used for the experiment.

In order to evaluate the estimation performance of BART at different traffic aggregation and time resolution for a large number of different  $Q$ , the experiment was carried out in phases. First, we studied two traffic cases of different aggregation level in a lab network with two Extreme Summit routers (Fig. 1). In each traffic case, the true cross traffic was recorded using the standard tool `tcpdump`, while the probe traffic receiver host recorded the series of measurements  $z_k$  and  $R_k$  for  $k = 1 \dots 2000$  (i.e. once every second for 2000 seconds). See below for details of traffic cases and BART probing setup.

Then, the true available bandwidth  $B_k$  for  $k = 1 \dots 2000$  was calculated from the cross traffic traces using a sliding window at eight different averaging time scales,  $\tau = 1, 2, 4, 8, 16, 32, 64$ , and 128 seconds. The tight link capacity was known due to the setup.

Finally, the BART filter estimation engine was fed the series of measurements  $(z_k, R_k)$  and produced the series of estimates  $\hat{B}_k$  for  $k = 1 \dots 2000$ . This was done for a wide range of  $Q$ . For each time scale  $\tau$ , these estimates were compared to the true available bandwidth  $B_k$  for  $k = 1 \dots 2000$ , and the optimal  $Q$  was identified.

As a metric for the precision of the estimation during the whole 2000 second tracking, we used the root mean square

of the relative estimation error (RMSE):

$$RMSE = \sqrt{\frac{1}{2000} \sum_{k=1}^{2000} \left( \frac{\hat{B}_k - B_k}{B_k} \right)^2} \quad (18)$$

In order to be able to tune the temporal characteristics of the estimation, we need to quantify the deviation from a straight line of the bandwidth as a function of time. As a metric for this, we introduce the quantity *variability*:

$$Variability = \sqrt{V\{\Delta B_k\}} = \sqrt{V\{B_2 - B_1, B_3 - B_2, \dots\}} \quad (19)$$

i.e. we use the standard deviation of the set of consecutive differences of bandwidth values.

**Traffic cases:** Synthetic cross traffic was generated, emulating a varying number of users in two different aggregation levels. In the two traffic cases used, the expectation value of the number of users was 10 and 100, respectively. New users arrive according to a Poisson process and remain active according to a Pareto distribution (shape = 1.5, mean = 1.0 second). When users are active, UDP packets are transmitted<sup>2</sup> with inter-arrival times following a Pareto distribution (shape = 1.9). The distribution parameters were chosen such that in both cases the expectation value of the overall cross traffic intensity was 5 Mbit/s.

BART was configured to transmit sequences of pairs of 1500-byte probe packets, organized as trains of 17 packets (16 pairs). The traffic intensity for each probe train was randomly chosen using a uniform distribution from 1 Mbit/s to 20 Mbit/s (more useful information is fed to the filter when we also probe at rates higher than the bottleneck capacity). The inter-departure time between two consecutive probe trains was set to one second; the average probe traffic overhead became 0.204 Mbit/s.

#### IV. RESULTS

Regarding the RMSE precision of the estimation, it soon became apparent that there is not much to be gained by tuning  $Q_{12}$ . This is illustrated in Fig. 2, which shows an example where the RMSE is plotted against  $Q_{12}$ , when  $Q_{11}$  and  $Q_{22}$  are fixed at their optimal values. The same behavior is mirrored for negative  $Q_{12}$ . It is clearly seen that the choice of  $Q_{12}$  is unimportant, as long as it stays within the allowed region (cf. section II C). For this reason, and for simplicity, we choose  $Q_{12}$  to be zero for the purposes of this paper. This leaves two independent elements of  $Q$  for tuning.

For each time scale  $\tau = 1, 2, 4, 8, 16, 32, 64,$  and 128 seconds, we systematically scanned the region  $10^{-6} \leq Q_{11} \leq 1, 10^{-6} \leq Q_{22} \leq 1$ , using logarithmically spaced grid points in both dimensions. Some of the results regarding the RMSE as a function of  $Q_{11}$  and  $Q_{22}$  are shown in Fig. 3-6 (one figure for each combination of 10 or 100 users and

averaging time scale  $\tau = 4$  seconds or 32 seconds).

We see in Fig. 3 and 4 that for 10 users,  $Q_{11} = 10^{-5}$  at the minimum for both time scales  $\tau = 4$  seconds and  $\tau = 32$  seconds. When comparing the time scales, the quality of the estimation is much more sensitive to  $Q_{22}$  than  $Q_{11}$ . The optimal  $Q_{22}$  is larger, and decreases as the time resolution increases. This is consistent with what is seen for the other  $\tau$ . In general, a low value of  $Q_{11}$  is preferred, together with a  $Q_{22}$  which decreases as  $\tau$  increases.

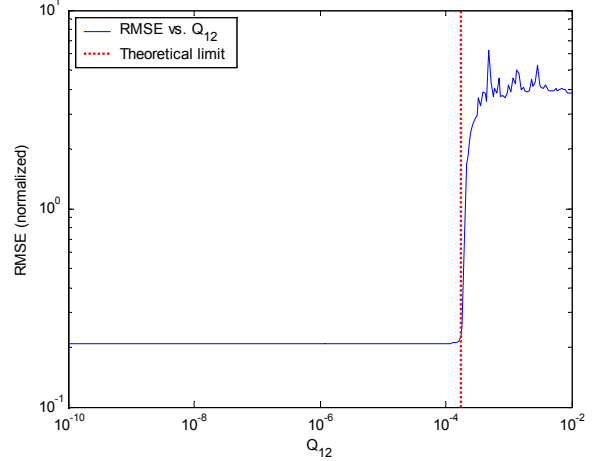


Fig. 2. RMSE when varying  $Q_{12}$ . The averaging time scale  $\tau$  is 4 seconds, applied to the 10 user traffic case.  $Q_{11}$  and  $Q_{22}$  have optimal values.

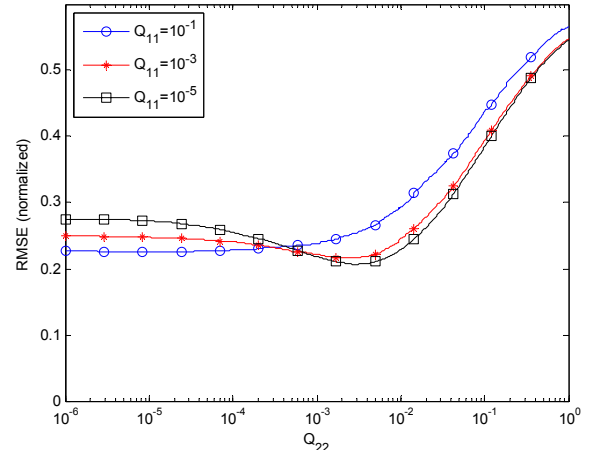


Fig. 3. RMSE for different  $Q_{11}$  and  $Q_{22}$  ( $Q_{12} = 0$ ). 10 users,  $\tau = 4$  seconds.

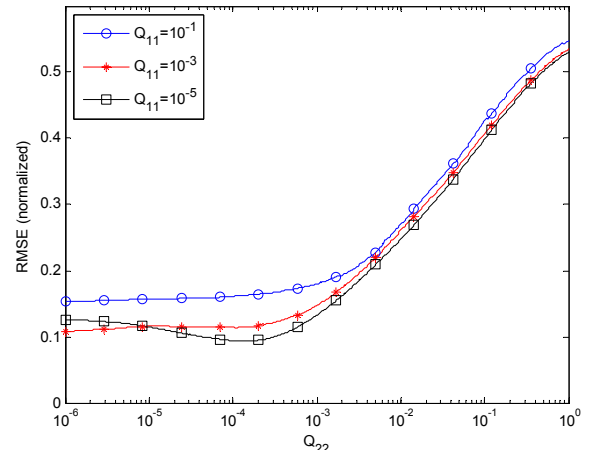


Fig. 4. RMSE for different  $Q_{11}$  and  $Q_{22}$  ( $Q_{12} = 0$ ). 10 users,  $\tau = 32$  seconds.

<sup>2</sup> The distribution of the synthetic cross traffic packet sizes roughly corresponds to observations from Sprint: <http://ipmon.sprintlabs.com/packstat/packetoverview.php> (December 2005)

For the high-aggregation traffic case of 100 users, we see in Fig. 5 and 6 that a good fit is achieved using low values of  $Q_{11}$  and  $Q_{22}$ . There is no clear minimum, but we see also that there is no apparent reason to choose  $Q_{11}$  different from  $10^{-5}$ , which turned out to be a suitable choice in the 10 user traffic case. However, a low RMSE is achieved as long as  $Q_{22} \leq 10^{-4}$ .

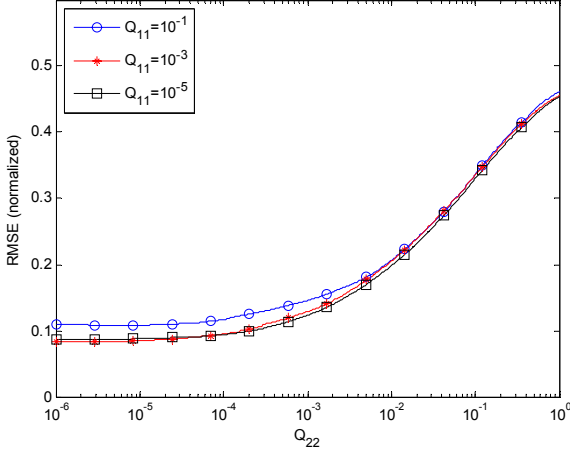


Fig. 5. RMSE for different  $Q_{11}$  and  $Q_{22}$  ( $Q_{12} = 0$ ). 100 users,  $\tau = 4$  seconds.

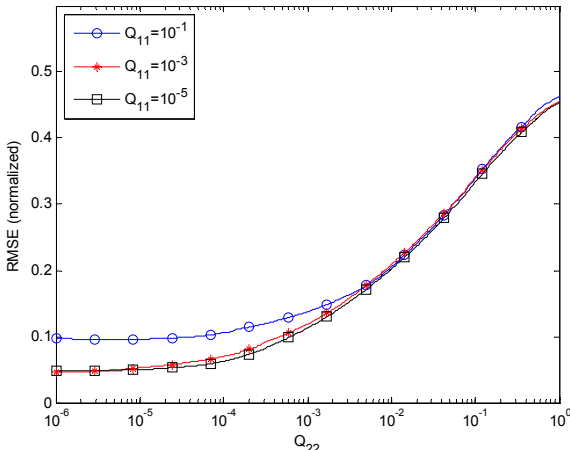


Fig. 6. RMSE for different  $Q_{11}$  and  $Q_{22}$  ( $Q_{12} = 0$ ). 100 users,  $\tau = 32$  seconds.

When we see that we can achieve good first-order characteristics of the estimation, i.e. good precision as measured by the RMSE, a natural next step is to go on to study the feasibility of reproducing higher-order characteristics, such as the temporal structure. We measure this using the variability, as defined in section III.

In Fig. 7 we see the variability of the available bandwidth versus  $Q_{22}$  for the traffic case corresponding to 10 users. The horizontal lines show the variability of the true available bandwidth at four different time resolutions. Of course, the true bandwidth does not depend on the estimation parameters; this is why these curves are horizontal lines. However, the variability depends on the time resolution  $\tau$ . The more fine-grained the time resolution, the higher is the variability. The other curves show the variability of the BART estimate as a function of  $Q_{22}$  for three different  $Q_{11}$ . We see that provided  $Q_{11}$  is chosen low enough, it is possible to choose  $Q_{22}$  such that the variability of the estimate matches the true variability of the available

bandwidth, for the desired time scales. This means that we may tune the method to reproduce the temporal characteristics of the true available bandwidth, by appropriately choosing  $Q_{22}$ .

Fig. 8 shows the same as Fig. 7, but for the traffic case corresponding to 100 users.

With respect to Fig. 7-8,  $Q_{11} = 10^{-5}$  turns out to be a suitable choice, since this makes it possible to restrict the variability tuning to  $Q_{22}$  and still obtaining a variability of the BART estimate that equals the true available bandwidth for a variety of time scales and cross traffic aggregation levels. For this reason, we choose  $Q_{11} = 10^{-5}$  for the remainder of this paper; from now on, the analysis concerns the correlation between  $Q_{22}$  and the cross traffic variability (i.e. the variability of the true available bandwidth).

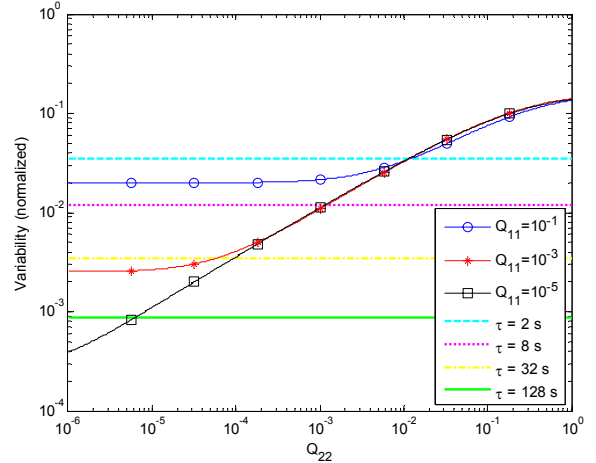


Fig. 7. Variability of the available bandwidth versus  $Q_{22}$  for 10 users.

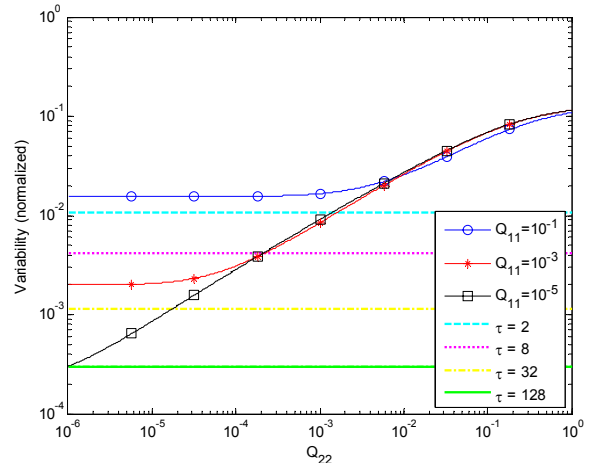


Fig. 8. Variability of the available bandwidth versus  $Q_{22}$  for 100 users.

Based on the variability, Fig. 9 depicts the optimum  $Q_{22}$  versus the time resolution. We observe a power-law behavior, but with clearly different proportionality constants for the different aggregation levels.

In Fig. 10, the variability of the true available bandwidth is plotted versus the time resolution. Again, a power-law behavior can be observed, but with clearly different proportionality constants for the different aggregation levels.

An interesting view of the correlation between the optimal  $Q_{22}$  and the variability of the cross traffic is offered by Fig. 11. Here, all 16 data points from the two different

traffic aggregation cases and the 8 different time scales are plotted for optimum  $Q_{22}$  versus the variability of the true available bandwidth. We can observe something rather close to a scaling behavior, in that the optimum  $Q_{22}$  exhibits a power-law dependence on the variability, while only showing a very weak dependence on the aggregation level.

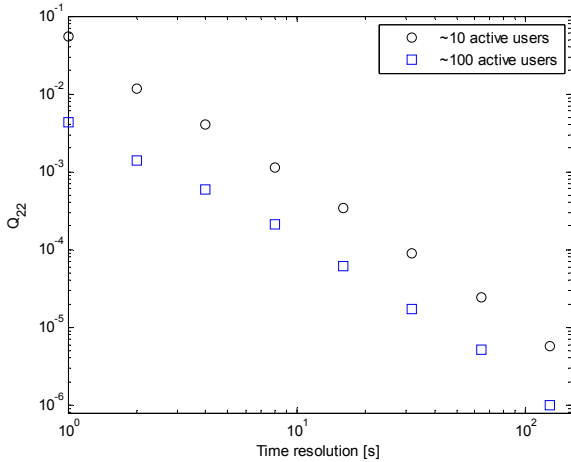


Fig. 9. Optimum  $Q_{22}$  versus time resolution ( $Q_{11} = 10^{-5}$ ,  $Q_{12} = 0$ ).

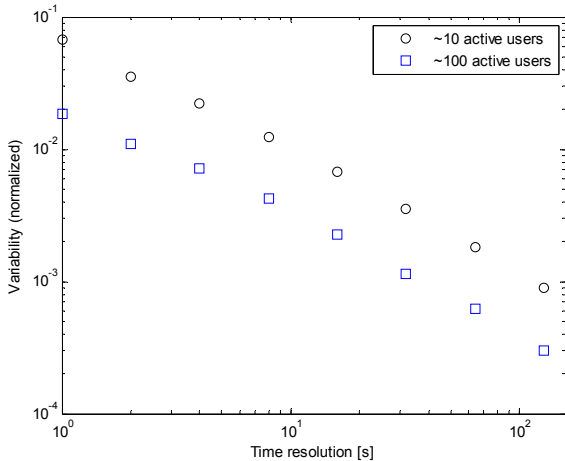


Fig. 10. Variability of the true available bandwidth versus time resolution.

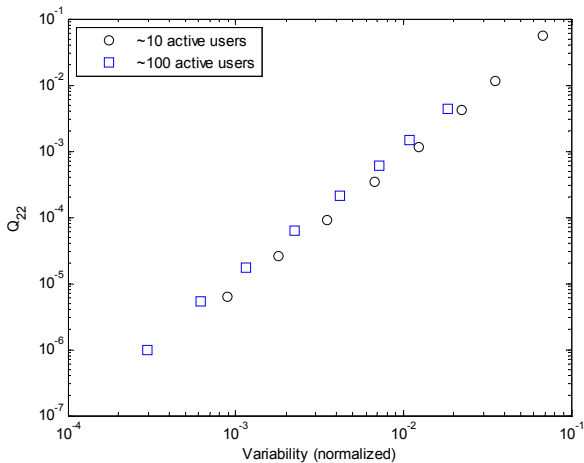


Fig. 11. Optimum  $Q_{22}$  versus variability of the true available bandwidth ( $Q_{11} = 10^{-5}$ ,  $Q_{12} = 0$ ).

Note that we expect the variability to be closely related to the input parameter  $Q$ , and in particular  $Q_{22}$ , of BART. From (15), remembering that  $\alpha$  is mostly constant, we see

that  $\Delta B$  and  $\Delta\beta$  are basically proportional. This means that the variability (see (19)) should essentially be proportional to the square root of  $Q_{22}$ , since from (17) we have  $Q_{22} = V(\Delta\beta)$ . This is precisely what we observe empirically in Fig. 11.

In Fig. 12-13, the resulting BART tracking of the available bandwidth in the 10 user traffic case can be compared to the true available bandwidth for the averaging time scales  $\tau = 4$  seconds and  $\tau = 32$  seconds, respectively. Note that the true available bandwidth is exactly the same in both figures, it is only averaged over different time resolutions. The only difference between the BART estimation algorithms in the two cases ( $\tau = 4$  seconds and  $\tau = 32$  seconds) is that  $Q_{22}$  has been set to the optimal value for each time scale, with respect to the variability. In Fig. 14-15, the same is displayed in the 100 user traffic case.

In Fig. 12-15, the BART estimate and the true available bandwidth are presented using physical units; i.e. no normalization.

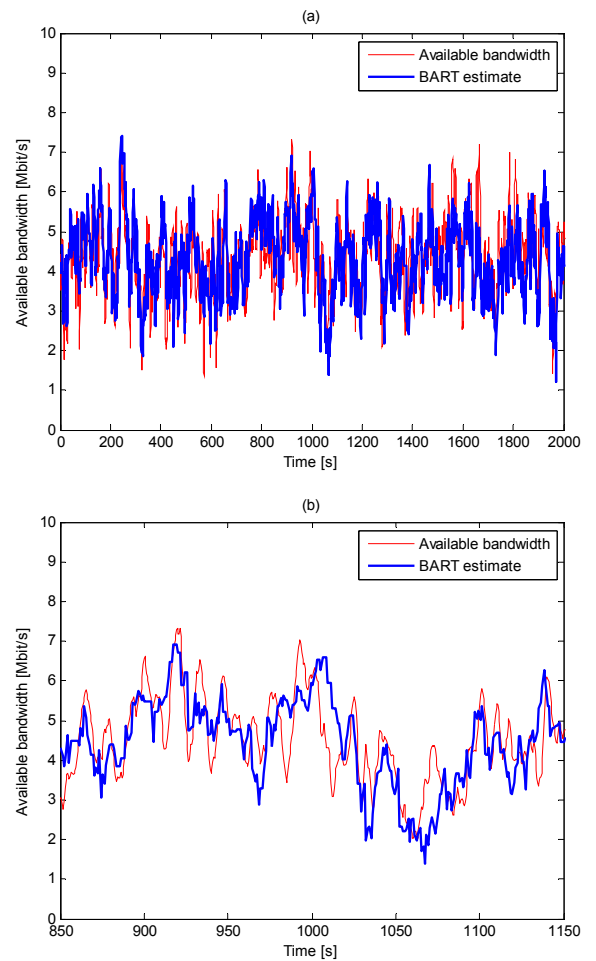


Fig. 12. (a) shows the true available bandwidth and the BART estimate at 4 seconds time resolution, 10 users. (b) shows the same, but zooming in at an arbitrary interval of 300 seconds.

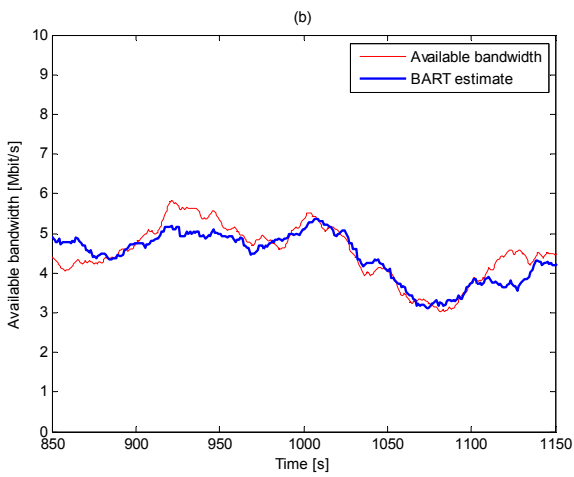
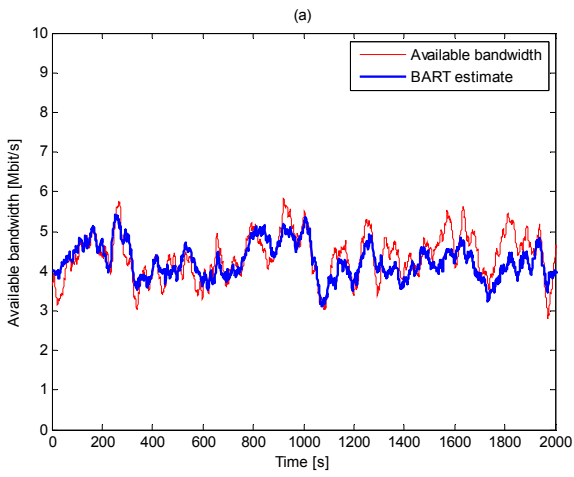


Fig. 13. (a) shows the true available bandwidth and the BART estimate at 32 seconds time resolution, 10 users. (b) shows the same, but zooming in at an arbitrary interval of 300 seconds.

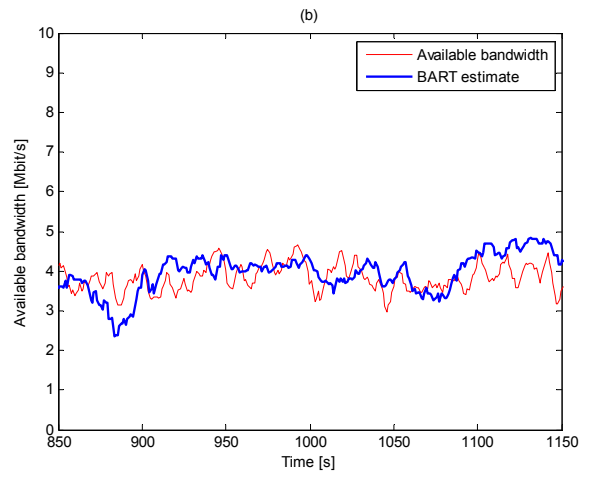
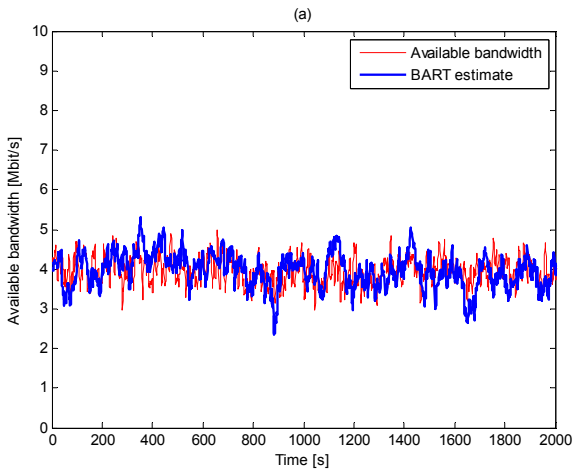


Fig. 14. (a) shows the true available bandwidth and the BART estimate at 4 seconds time resolution, 100 users. (b) shows the same, but zooming in at an arbitrary interval of 300 seconds.

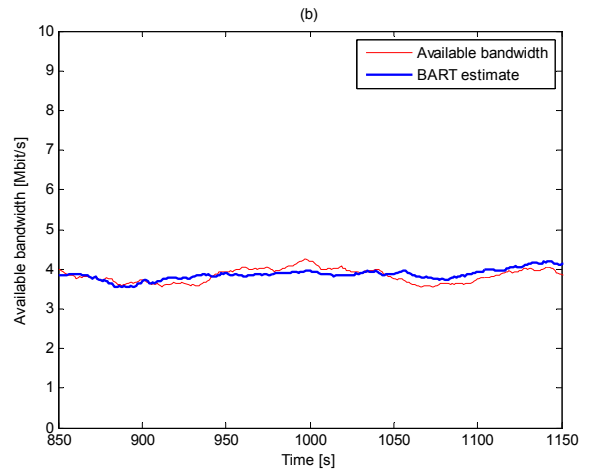
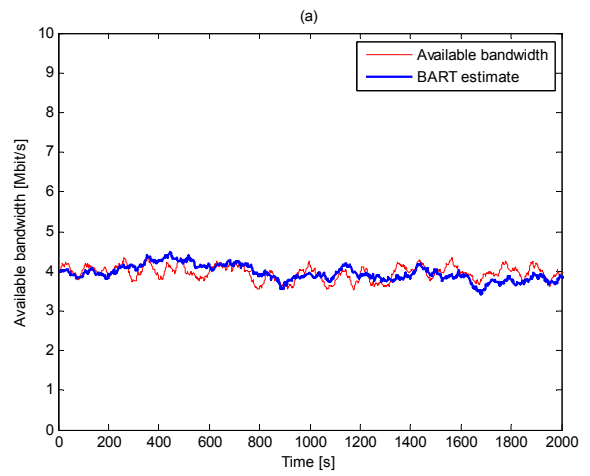


Fig. 15. (a) shows the true available bandwidth and the BART estimate at 32 seconds time resolution, 100 users. (b) shows the same, but zooming in at an arbitrary interval of 300 seconds.

We have studied a tuning opportunity of the BART Kalman filter-based method for bandwidth estimation, namely the tuning of the three free elements of the symmetric  $2 \times 2$  matrix  $Q$  for optimizing the tracking quality for desired temporal characteristics.

First, we made use of the RMSE metric in order to investigate the precision of the BART estimate. It was found that when  $Q_{12}$  was chosen outside its theoretical bounds, the estimation quality broke down (cf. Fig. 2). This can be interpreted as a “sanity check” of the method. Further, we found no performance gain when allowing nonzero values; thus, we chose to fix  $Q_{12} = 0$ .

Regarding  $Q_{11}$  and  $Q_{22}$ , we found that the RMSE is highly dependent upon  $Q_{22}$ , whereas the choice of  $Q_{11}$  is less critical. The optimal value of  $Q_{22}$  was observed to decrease with increasing bandwidth averaging time scale  $\tau$  and with increasing cross traffic aggregation.

A specific variability measure was suggested and used, in order to be able to tune the temporal characteristics of the estimation. It turned out that as long as  $Q_{11}$  is small enough,  $Q_{22}$  is the crucial element when tuning  $Q$  for bandwidth estimation variability. The recommendation is to assign  $Q_{11}$  a constant and small value, say  $Q_{11} = 10^{-5}$  (note that this suggestion does not conflict with the results obtained from the RMSE investigation). Setting  $Q_{11}$  exactly equal to zero does not work well. The behavior of the filter would be too dependent on the choice of initial values  $x_0$  and  $P_0$ . Also, the filter would probably perform poorly in the case of a change in the state variable  $\alpha$ , i.e. the bottleneck link capacity. Consequently, it is better to allow for some uncertainty by choosing a small nonzero value for  $Q_{11}$ .

Provided  $Q_{11}$  is given a small value, it is possible to obtain a large range of variabilities of the BART estimate by simply tuning  $Q_{22}$ . The optimal  $Q_{22}$  decreases as the variability of the true available bandwidth decreases (i.e.  $Q_{22}$  decreases with increasing time scale and cross traffic aggregation). Also, from Fig. 11, it appears that, when fitted to a power-law behavior, the optimal  $Q_{22}$  roughly scales with the variability, with an exponent of 2.

Overall, the experimental results are consistent with what might be expected:

- $Q_{11}$  and  $Q_{12}$  should be small (cf. the discussion in section II C).
- For rapid response to new measurements, which is needed for tracking at high variability,  $Q_{22}$  should be large. For a more stable tracking, suitable for low variability,  $Q_{22}$  should be small (cf. the argument in the end of section II A).

In conclusion, we have found that it is feasible to tune the BART filter-based estimation method for enhanced tracking performance at desired available bandwidth variability, by choosing the input parameter  $Q$  appropriately.

- [1] S. Ekelin, M. Nilsson, E. Hartikainen, A. Johnsson, J.-E. Mångs, B. Melander, and M. Björkman, “Real-time measurement of end-to-end available bandwidth using Kalman filtering,” in *Proc. 10th IEEE/IFIP Network Operations and Management Symposium*, Vancouver, Canada, 2006.
- [2] R. Prasad, M. Murray, C. Dovrolis, K. Claffy, “Bandwidth estimation: metrics, measurement techniques, and tools,” in *IEEE Network Magazine*, November/December, 2003.
- [3] J. Navratil and R.L. Cottrell, “Abwe: a practical approach to available bandwidth estimation,” in *Proc. Passive and Active Measurement Workshop*, 2003.
- [4] R. Carter and M. Crovella, “Measuring bottleneck link speed in packet-switched networks,” in Technical Report 96-006, Boston University, 1996.
- [5] N. Hu and P. Steenkiste, “Evaluation and characterization of available bandwidth probing techniques,” in *Proc. IEEE JSAC Internet and WWW Measurement, Mapping, and Modeling*, 2003.
- [6] V. Ribeiro, R. Riedi, G. Baraniuk, J. Navratil, and L. Cottrell, “pathChirp: efficient available bandwidth estimation for network paths,” in *Proc. Passive and Active Measurement Workshop*, 2003.
- [7] G. Jin, G. Yang, B.R. Crowley, and D.A. Agarwal, “Network characterization service (NCS),” in Lawrence Berkely National Lab Report 47892, 2001.
- [8] M. Jain and C. Dovrolis, “Pathload: a measurement tool for end-to-end available bandwidth,” in *Proc. Passive and Active Measurement Workshop*, 2002.
- [9] J. Strauss, D. Katabi, and F. Kaashoek, “A measurement study of available bandwidth estimation tools,” in *Proc. ACM SIGCOMM Internet Measurement Conference*, 2003.
- [10] B. Melander, M. Björkman, P. Gunningberg, “A new end-to-end probing and analysis method for estimating bandwidth bottlenecks,” in *Proc. IEEE Globecom '00*, San Francisco, USA, 2000.
- [11] F. Montesino-Pouzols, “Comparative analysis of active bandwidth estimation tools,” in *Proc. Passive and Active Measurement Workshop*, 2004.
- [12] A. Shriram, M. Murray, Y. Hyun, N. Brownlee, A. Broido, M. Fomenkov, K. Claffy, “Comparison of public end-to-end bandwidth estimation tools on high-speed links,” in *Proc. Passive and Active Measurement Workshop*, 2005.
- [13] S. Keshav, “A control-theoretic approach to flow control,” in *Proc. ACM SIGCOMM '91*, 1991.
- [14] G. Bishop and G. Welch, “An introduction to the Kalman filter,” in *SIGGRAPH 2001*, Course 8, 2001.