

Distributed Bandwidth Reservation by Probing for Available Bandwidth

Sven Krasser, Henry L. Owen
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0250, USA
Email: {sven, owen}@ece.gatech.edu

Jochen Grimminger, Hans-Peter Huth,
Joachim Sokol
Siemens AG, CT IC2 Corporate Technology
81730 Munich, Germany
Email: {jochen.grimminger, hans-peter.huth,
joachim.sokol}@mchp.siemens.de

Abstract—For many applications in Internet protocol-based networks a guaranteed quality of service is crucial. A simple and scalable way of achieving this is to use the differentiated services architecture in conjunction with admission control. In this research, we focus on the reservation of bandwidth by routers on the edge of a single network domain with differentiated services support. We evaluate the assumptions made by previous research on probe-based bandwidth reservation. The availability of bandwidth is determined by measuring the throughput a probe flow of a certain rate achieves while the reservation is accomplished by assuming that the throughput this probe flow achieves equals to an amount of reserved bandwidth.

Index Terms—Quality of Service, Call Admission Control, Available Bandwidth, Network Probing

I. INTRODUCTION

As the Internet evolves and other networks migrate to Internet technologies, the need for a guaranteed quality of service (QoS) in such Internetworks becomes more and more apparent. Certain services require firm constraints on e.g. bandwidth or delay. The integrated services (IntServ) architecture and the resource reservation protocol (RSVP) address these issues but they have shortcomings with regard to scalability. The differentiated services (DiffServ) architecture enables the network to optimize the transport of data packets according to certain requirements but it gives no guarantees on these transport characteristics. However, for e.g. telephone networks, there are firm constraints on delay and bandwidth that have to be adhered to for a connection of reasonable quality. In this paper, we focus on the bandwidth requirements.

There are several approaches to address these resource-related problems. In the Scalable Resource Reservation Protocol framework [1], a user injects packets in the network with a request bit set. The network forwards these packets at a rate it is going to accept. The receiver reports the rate these packets are received to the sender. Now the sender sets a reserved bit rather than the request bit set previously. The network in turn tries to forward these packets. Scalability is achieved since there is no per-flow information needed in the network.

Another approach is to use admission control to keep the drop rate in the network reasonably low. The admission control can be based on end-to-end measurements of packet transmission characteristics. The frameworks presented in [2], [3], and [4] use network probes for admission control. Before

a host sends regular data packets, it sends probes and measures the characteristics of this probe transmission. If these characteristics suffice some requirements, it is assumed that the network can support the additional traffic, and the host starts to send regular data packets. The latter receive a better service from the network than probes do by mapping them to a different queue.

This idea is further refined in the two-phase edge-to-edge distributed measurement based admission control (TPED MBAC) [5]. Probing is done in two steps. In the quantitative provisioning phase, it is determined if bandwidth requirements are met. In the qualitative provision phase it is evaluated if the QoS requirements are fulfilled.

In [6], a scheme is introduced that incorporates aspects from these frameworks to get an unintrusive mechanism for bandwidth reservation. The assumed network consists of core routers and edge routers. The latter have the responsibility to reserve bandwidth before using it. Edge routers are ingress and egress points for traffic.

This scheme is designed to fit call admission control (CAC) needs in radio access networks (RANs). Ingress routers for traffic have to check more than just whether the network can support a call during the initial call setup. Instead, since nodes can move, ingress routers have also to ensure that they can support a certain amount of traffic that is due to roaming. Mobile nodes can connect to a new ingress router after a call has been accepted by their initial ingress router. Therefore, to assure a fast transition, the new ingress router needs a fast means to determine whether it can support this call. This is achieved by continuous probing in an unintrusive fashion.

The rest of this paper is organized as follows. Section II gives a brief overview over the unintrusive probing scheme originally introduced in [6]. Section III addresses some of the assumptions of that mechanism and investigates the characteristics of the framework if not all of these assumptions hold. In Section IV we describe the actual bandwidth reservation mechanism. Section V concludes the paper.

II. UNINTRUSIVE BANDWIDTH PROBING

A. Overview

The basic idea of this probing scheme is to use DiffServ and a special probe per-hop behavior (PHB). This PHB is set

up such that routers forward probes only if there is no data packet available. Ingress routers send probes to egress routers at a rate that corresponds to their additional bandwidth needs. The egress routers periodically send back the observed rate the probes are received with to the ingress routers. Based on the rate the probes are sent with originally and the rate the probes are received (as reported by the egress router), the ingress router draws conclusions about the amount of bandwidth that is available and reserved for it.

B. Nomenclature and Assumptions

Probes are sent between edge routers, which we call E1 to E_{n_e} throughout this paper. The actual data is generated by nodes N1 to N_{n_n} connected to the network that is bordered by these edge routers. The remainder of the network consists of the core routers C1 to C_{n_c} . For reasons of simplicity we assume that all traffic injected into the core network by an edge router E_i is generated by node N_i . Hence, $n_e = n_n$. Edge router E_j sends probes to edge router E_k at a rate $R_{j,k}$. Note that in general there can be multiple paths connecting those routers. For this research, we assume a single path is used when E_j is sending to E_k . This does not affect our evaluation but simplifies the nomenclature since a path can be uniquely identified by the indices j and k of its edge routers. $D_{j,k}$ denotes the data rate N_j is sending to N_k . This implies under our assumptions that this traffic has E_j as ingress point into the core network, while E_k is the egress point. Furthermore, this traffic traverses the same path as the probes that E_j sends to E_k . E_k measures a rate $M_{j,k}$ of probes coming in from E_j .

C. Replacing Probes with Data Packets

In [6], probing over a single congestable link is evaluated. In that scenario, let the overall probe rate of all edge routers probing over this single congestable link be S and the available bandwidth not being used for data traffic be A . Additionally, let $A < S$. If a fluid traffic model is used, the measured rate $M_{j,k}$ can be determined as

$$M_{j,k} = R_{j,k} \cdot \frac{A}{S}. \quad (1)$$

If E_j wants to increase its data rate $D_{j,k}$ by $Q_{j,k}$, it has to decrease its probe rate $R_{j,k}$ by $Q_{j,k}^{probe} = Q_{j,k} \cdot \frac{R_{j,k}}{M_{j,k}}$ in order that other probe flows do not measure a lower rate. Before the rates change, E_m measures a probe rate $M_{l,m}$ originating from E_l according to

$$M_{l,m} = R_{l,m} \cdot \frac{A}{S}.$$

After the rates change, the new measured probe rate is

$$M_{l,m}^{new} = R_{l,m} \cdot \frac{A - Q_{j,k}}{S - Q_{j,k} \cdot \frac{R_{j,k}}{M_{j,k}}}.$$

By using (1) to substitute $M_{j,k}$, we obtain

$$M_{l,m}^{new} = R_{l,m} \cdot \frac{A - Q_{j,k}}{S - Q_{j,k} \cdot \frac{S}{A}} = R_{l,m} \cdot \frac{A}{S} = M_{l,m}.$$

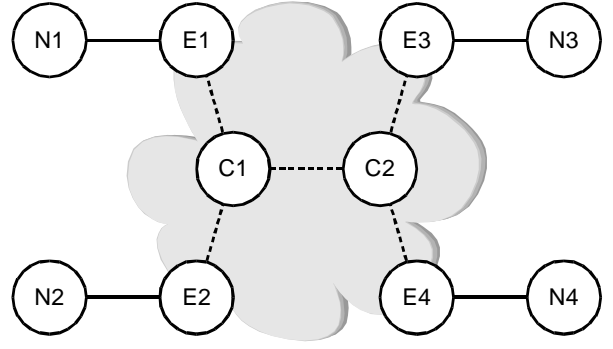


Fig. 1. Topology of simulated test network.

Since ingress routers can be sure that the probe rate they achieve does not change, this bandwidth can be seen as reserved. Ingress router E_j can increase the rate of the data it is injecting into the core network bound for egress router E_k by up to $Q_{j,k}^{max} = M_{j,k}$.

III. ASSUMPTIONS AND THEIR IMPACT

As already denoted in the previous section, some assumptions have been made previously for the purpose of a straight-forward evaluation. In general, not all of these assumptions hold or are tolerable in reality.

A. Accuracy of Measured Rates at Egress Routers

The calculation of the correct value the probe rate has to be decreased by relies on the exact value $M_{j,k}$. In reality, this value is measured by the egress router E_k , and the result is sent back to E_j . Obviously, the actual measured rate, which we denote as $\tilde{M}_{j,k}$, does in general differ from the theoretical value calculated in (1).

We present simulation results showing this effect for a simple topology. Fig. 1 shows this topology. The cloud denotes the probing domain with the inner core routers C1/C2 and edge routers E1 to E4. Initially, the data rates are configured as $R_{1,3} = 70$ kb/s, $D_{1,3} = 20$ kb/s, $R_{2,3} = R_{2,4} = 5$ kb/s, and $D_{2,3} = D_{2,4} = 5$ kb/s. The data rate $D_{1,3}$ is varied by a value $Q_{1,3} = -20 \dots 70$ kb/s ($D_{1,3}^{new} = D_{1,3} + Q_{1,3}$). Likewise, the probe rate is varied by $Q_{j,k}^{probe} = Q_{j,k} \cdot \frac{R_{1,3}}{\tilde{M}_{1,3}}$, where $\tilde{M}_{1,3}$ is set to the correct rate of $M_{1,3} = 52.5$ kb/s and to values 10% below or above this rate.

Fig. 2 shows the theoretically anticipated measured rate $\tilde{M}_{2,4}$ for the probe flow from E2 to E4 depending on the rates of the data and probe flows from N1/E1 to N3/E3. The figure depicts variations of $\tilde{M}_{2,4}$ as $D_{1,3}$ is increased by $Q_{1,3}$ and $R_{1,3}$ is decreased by $Q_{1,3}^{probe} = Q_{1,3} \cdot \frac{R_{1,3}}{\tilde{M}_{1,3}}$ (or vice versa, if $Q_{1,3}$ is negative and thus $D_{1,3}$ is decreased and $R_{1,3}$ is increased). Ideally, the measurement of $\tilde{M}_{1,3}$ equals $M_{1,3}$, the value predicted by (1) based on a fluid traffic model. Hence, $\tilde{M}_{1,3} = 52.5$ kb/s (before the rates from N1/E1 to N3/E3 are changed based on $Q_{1,3}$). In this case, if $D_{1,3}$ and $R_{1,3}$ are changed, then there is no effect on $\tilde{M}_{2,4}$ as long as the sum of all rates are below the bottleneck bandwidth. If the measurement is 10% too low and hence $\tilde{M}_{1,3} = 47.25$

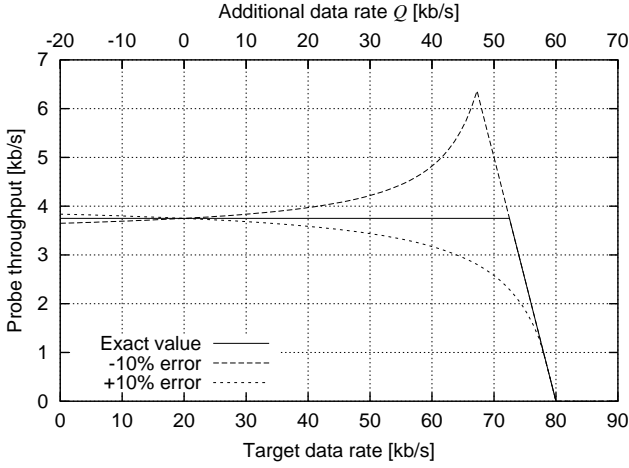


Fig. 2. Theoretically anticipated rate $\tilde{M}_{2,4}$.

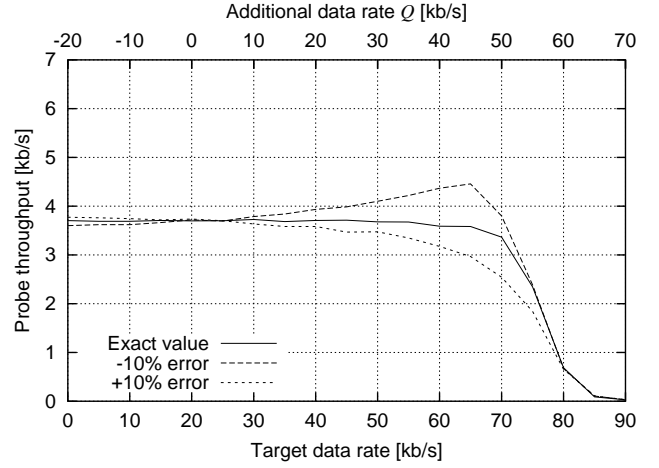


Fig. 3. Simulated results for $\tilde{M}_{2,4}$.

kb/s, $\tilde{M}_{2,4}$ increases as more probes are replaced with data. In this erroneous measurement case all probes are replaced for $Q_{1,3} = 47.25$ kb/s. Then, $\tilde{M}_{2,4}$ linearly decreases as $D_{1,3}$ is increased further. In the case of a 10% too high measurement, $\tilde{M}_{1,3} = 57.75$ kb/s, $\tilde{M}_{2,4}$ decreases since $R_{1,3}$ is not decreased enough. When all probes are replaced ($Q_{1,3} = 57.75$ kb/s), the curve again decreases linearly.

Fig. 3 shows the experimental results obtained by simulation. The simulated rates are close to the mathematically estimated ones in Fig. 2. Note that these results represent the effects when one flow measures an increased rate while the other flows do not measure a decreased rate. In reality, if one flow measures an increased rate, other flows measure a decreased rate. Hence, these nodes assume a lower value for their reserved bandwidth anyway. There are however some unusual situations when an egress router measures an increased rate, but the other routers do not measure a decreased rate, e.g. if lots of probes are piling up in a queue near the egress router and are then transmitted in a burst.

The setup analyzed here consists of only three probe flows. Therefore, the probe throughput degradation experienced by the flows on paths for which the data rate is not increased is more severe than in a situation where there are lots of other flows. Hence, we conclude that there is no significant effect of measurement errors on bandwidth reservation in general.

B. Variable Bit Rate Traffic

The bandwidth reservation mechanism assumes that once the probe rate is reduced, data is sent at a constant bit rate (CBR). In general, the traffic characteristics of many applications show a variable bit rate (VBR).

VBR issues for the probing framework presented in [4] are presented in [7]. An important observation in the latter paper is that one cannot use the average rate to reduce the VBR to a CBR case. If multiple VBR sources peak, the network becomes overloaded. Instead, the peak data rate of flows has to be considered. Fig. 4 shows two VBR data flows sharing bandwidth. Although the sum of both average rates is in

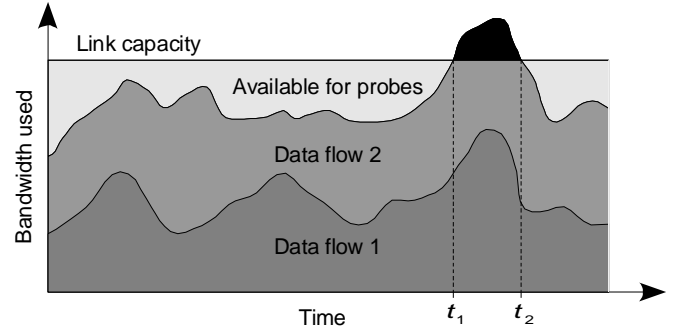


Fig. 4. Example bandwidth utilization for two data flows.

general lower than the total bandwidth available, the network is overloaded between t_1 and t_2 .

In [7], it is proposed to use a sufficiently long probing phase to measure all VBR characteristics before a call is admitted. In contrast to [4], the probing scheme in this research continuously probes the network rather than using probing phases and data phases for distinct calls. Hence, VBR traffic can be addressed by applying a filter function on the measured probe rate at the egress router. Since the probed rate is not subject to changes, the implementation of an appropriate filter function is straightforward.

Another solution is to monitor the data rate. If this rate drops, the ingress router can fall back to probes again. The latter has the disadvantage that there is no clear instantaneous rate. Rates are always averaged over a certain amount of time. Therefore, it is hard to decide when the data rate drops.

C. Multiple Bottlenecks

The fluid traffic model used to calculate the rate Q^{probe} that probing has to decreased by assumes a single bottleneck. In Fig. 5 we show a probing scenario over multiple bottlenecks with the corresponding nomenclature. On link 0, probes are sent with a rate R_0 . Due to cross traffic (denoted by a gray arrow) on link 1, the total probing rate is S_1 , and the free bandwidth that can be used by probes is A_1 . Moreover, we

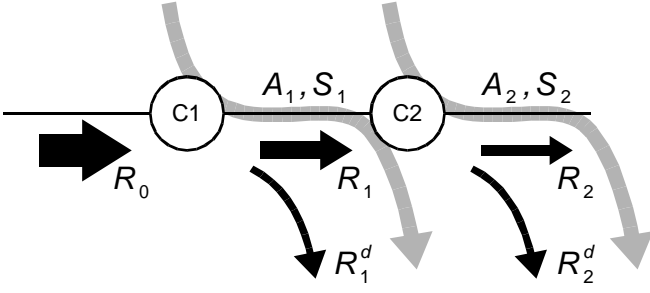


Fig. 5. Probing through multiple bottlenecks.

assume that $A_i < S_i$ for all bottleneck links i . Because of these constraints, core router C1 forwards probes at a decreased rate of $R_1 = R_0 \cdot \frac{A_1}{S_1}$. Probes are dropped at a rate of $R_1^d = R_0 - R_1$ at C1. A similar situation occurs on the next link, link 2. The probe drop rate is $R_2^d = R_1 - R_2$ while probes are forwarded by C2 with a rate $R_2 = R_1 \cdot \frac{A_2}{S_2}$. The available bandwidth for probes on this link is A_2 , the overall probe rate is S_2 . Both link 1 and link 2 are congested. Hence,

$$R_0 > R_1 > R_2. \quad (2)$$

Eventually, R_2 is measured at an egress router as the rate M_0 corresponding to the original rate R_0 . Based on this, Q_{probe} is calculated as

$$Q_{probe} = Q \cdot \frac{R_0}{M_0} = Q \cdot \frac{R_0}{R_2} = Q \cdot \frac{S_1 S_2}{A_1 A_2}. \quad (3)$$

Note that S_1 and S_2 depend on R_1 and on R_2 , respectively, since S_i denotes the sum of all probe rates on link i .

However, Q_{probe} is based on a single bottleneck assumption and hence is not the correct rate. This value of Q_{probe} does not assure that other flows are not affected. Additionally, every link would require a different value for Q_{probe} so that other flows do not measure altered rates. For link 1, this value is

$$Q_1^{probe} = Q \cdot \frac{R_0}{R_1}. \quad (4)$$

For link 2 we get

$$Q_2^{probe} = Q \cdot \frac{R_1}{R_2}. \quad (5)$$

By using (2), we can conclude from (3), (4), and (5) that

$$Q_{probe} = Q \cdot \frac{R_0}{R_2} > Q \cdot \frac{R_0}{R_1} = Q_1^{probe}$$

and

$$Q_{probe} = Q \cdot \frac{R_0}{R_2} > Q \cdot \frac{R_1}{R_2} = Q_2^{probe}.$$

This argument still holds in the general case of more than two bottlenecks.

Therefore, if a single bottleneck is assumed to calculate Q_{probe} , but there are multiple bottlenecks present in the network, the ingress router will back off its probe rate more than actually necessary. In other words, it has more bandwidth available than measured. The effect on other probe flows is that they measure that more bandwidth is reserved for them after

the rates have been changed by the ingress router of interest. This is not severe since every ingress router can still assume that the amount of bandwidth reported by a corresponding egress router is still available. More precisely, in a multiple bottleneck setup ingress routers can assume that *at least* the bandwidth reported to them is available.

IV. BANDWIDTH RESERVATION

The framework presented so far deals with the effects of replacing a given amount of the probe rate with data. It does not, however, explain how this probe rate is chosen by ingress routers in the first place and how the reserved bandwidth can be increased or decreased.

In this section we outline two approaches that we are currently investigating. Both rely on the results presented in Section II C, i.e. that data and probe rates can be changed without changing the probe rates others observe.

A. Uncooperative Scheme

The first scheme requires nearly no cooperation between edge routers. With respect to probing, two distinct cases of link loads are relevant. First, all data rates plus all probe rates are less or equal than the link capacity C . This implies that $A \geq S$. Second, all data rates are less than C , but all data rates plus all probe rates are more than it. Therefore, $A < S$. Fig. 6 shows these values for a single link partially used by data and probe traffic.

If $A \geq S$, ideally no probe is dropped. Probes can be dropped occasionally because of packet bursts temporarily overloading a queue. Hence, $M_{j,k} = R_{j,k}$. Let B denote the bandwidth completely unused. Then an ingress router can increase its data rate $D_{j,k}$ by $Q_{j,k}$ without decreasing its probe rate $R_{j,k}$ as long as $Q_{j,k} \leq B$. In general, if multiple ingress routers increase their data rates, the sum of these increases has to be either less than or equal to B (assuming one congestable link). The same holds for increasing the probe rate. However, B is not known by the ingress routers. Routers can determine whether B is exceeded by observing if the measured rate $M_{j,k}$ is smaller than $R_{j,k}$. Since the actual measured rate $\tilde{M}_{j,k}$ is in general slightly inaccurate, the routers should determine whether B is exceeded with the help of a tolerance factor α . Hence, B is exceeded if $\tilde{M}_{j,k} < \alpha \cdot R_{j,k}$. Since multiple routers can increase their rates at the same time and exceed B , the measured probe rate can drop spontaneously leading to less reserved bandwidth. Therefore, if B is assumed not to be exceeded, routers take care for this effect by only claiming a fraction β of this rate as reserved. The assumed reserved bandwidth for E_j sending to E_k is $\min(\tilde{M}_{j,k}, \beta \cdot R_{j,k})$. Eventually, all capacity will be used, and when routers increase their data rate, they decrease their probe rate according to Section II C.

B. Cooperative Scheme

The second scheme assumes that edge routers cooperate with respect to reservations although it sticks with the paradigm that no direct signaling (besides feedback packets) is

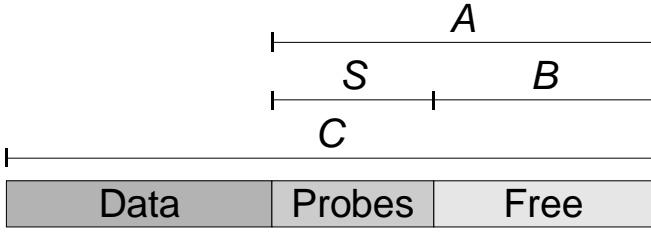


Fig. 6. Rates and capacities.

exchanged among edge routers. For reasons of simplicity, we assume a single congestable link. In this setup, ingress routers can be sure that the probe rate received at the egress router is constant in the long run although there can be instantaneous fluctuations that are due to VBR traffic. Moreover, if we drop the single congestable link assumption, the probe rate can only be increased in the long run while it will never be decreased. Therefore, an edge router can increase its rates in a way that other routers receive probes at a decreased rate. These routers then assume that the decreased rate is due to such an immoderate rate increase of another edge router. Likewise, they still assume that the bandwidth reserved is the probe rate received previously. Additionally, if this decreased rate is received for a certain amount of time t_d , it is interpreted as a request to withdraw some of the reserved bandwidth because it is needed by another edge router. To make this work, the decrease has to be big enough such that edge routers can distinguish it from bursts induced by VBR traffic. For the edge router E_j probing to E_k , a way to achieve the rate decrease other edge routers measure is to increase its own probe rate by $P_{j,k}$. Then other edge routers, say probing from E_l to E_m , will not measure the normal rate of

$$M_{l,m} = R_{l,m} \cdot \frac{A}{S}.$$

They measure a decreased rate of

$$M_{l,m}^{new,P} = R_{l,m} \cdot \frac{A}{S + P_{j,k}}.$$

Since probes are dropped on their way through the network, the rate of $P_{j,k}$ remaining can become too small to have an impact on $M_{l,m}^{new,P}$. A remedy to this is the use of a privileged probe type. These privileged probes have a higher priority than regular probes but a lower priority than data packets. By default, edge routers send no privileged probes. If a decrease of other probe flows is required from E_j to E_k , edge router E_j sends out these privileged probes at a rate $H_{j,k}$ to E_k . This has basically an effect on A . Finally, other edge routers measure

$$M_{l,m}^{new,P,H} = R_{l,m} \cdot \frac{A - H_{j,k}}{S + P_{j,k}}.$$

Lets turn back to the probe rates from E_j to E_k , still assuming a single bottleneck. Basically, we can distinguish four steps. First, the normal rate probes are received with is given by

$$M_{j,k} = R_{j,k} \cdot \frac{A}{S}. \quad (6)$$

Then E_j increases $R_{j,k}$ by $P_{j,k}$. The received probe rate changes to

$$M_{j,k}^{new,P} = (R_{j,k} + P_{j,k}) \cdot \frac{A}{S + P_{j,k}}. \quad (7)$$

By using (6) and (7), we can estimate A as

$$A = \frac{M_{j,k} M_{j,k}^{new,P} P_{j,k}}{M_{j,k} R_{j,k} + M_{j,k} P_{j,k} - R_{j,k} M_{j,k}^{new,P}}.$$

Based on this we choose a value $H_{j,k} < A$. Third, we send out privileged probes at this rate. The rate received is

$$M_{j,k}^{new,P,H} = (R_{j,k} + P_{j,k}) \cdot \frac{A - H_{j,k}}{S + P_{j,k}}.$$

Eventually, other nodes will back off and release parts of the bandwidth they reserved. This amount is denoted by T . The rate received after this step is

$$M_{j,k}^{new,P,H,T} = (R_{j,k} + P_{j,k}) \cdot \frac{A - H_{j,k}}{S - T + P_{j,k}}.$$

T can now be estimated as

$$T = \frac{(M_{j,k} P_{j,k} + R_{j,k} H_{j,k})(M_{j,k}^{new,P,H} - M_{j,k}^{new,P,H,T})}{M_{j,k}^{new,P,H,T}(M_{j,k} - M_{j,k}^{new,P,H})}.$$

This amount of bandwidth can then be reserved by E_j by setting the new probe rate to $R_{j,k}^{new} = R_{j,k} + T$. However, this analysis only works for single bottlenecks. In a multiple bottleneck setup, T cannot be estimated by using this analysis.

V. CONCLUSION

We presented a probing mechanism that guarantees the availability of bandwidth. There is no need for long setup phases if an ingress router wants to admit traffic fast, so that roaming mobile nodes are supported. If the network gets overloaded, this mechanism guarantees a fair share of the remaining resources in the network.

Building on this mechanism, we presented outlines for reservation schemes that rely on congesting the links with probes. By congesting the links, information about other probe flows can be gathered yielding a distributed reservation mechanism. Such information can also be used for indirect signaling, as we have shown in Section IV B. However, congesting the links with probes also implies that bandwidth used for best effort traffic has also to be reserved. We are also researching probing/reservation schemes that do not rely on this congestion assumption. This can be achieved by other two-level probing mechanisms similar to the one presented in Section IV B. Bandwidth reservation probes always achieve their initial rate ($R = M$) while the remaining bandwidth (B) is probed with lower priority probes to determine the availability of bandwidth.

Another issue is the way multiple bottlenecks can be accounted for. Especially the analysis in Section IV B shows that the single bottleneck assumption works well, but that not all results generalize to multiple bottlenecks like we demonstrated in Section III C. Additionally, it is difficult to gather a

sufficient number of parameters by using end-to-end probing techniques in such a setup. We are currently investigating multiple bottleneck setups and parameter estimation in such scenarios.

We plan to incorporate a bandwidth reservation/probing scheme like this in a radio access network and evaluate other probed metrics like delay, jitter, and packet loss.

REFERENCES

- [1] W. Almesberger, T. Ferrari, and J.-Y. Le Boudec, "SRP: a scalable resource reservation protocol for the Internet," in *Proc. IEEE IWQoS '98*, vol. 1b, 1998, pp. 831–836.
- [2] V. Elek, G. Karlsson, and R. Ronngren, "Admission control based on end-to-end measurements," in *Proc. IEEE INFOCOM 2000*, 2000, pp. 623–630.
- [3] G. Bianchi, A. Capone, and C. Petrioli, "Throughput analysis of end-to-end measurement-based admission control in IP," in *Proc. IEEE INFOCOM 2000*, vol. 3, 2000, pp. 1461–1470.
- [4] F. Borgonovo, A. Capone, L. Fratta, M. Marchese, and C. Petrioli, "PCP: A bandwidth guaranteed transport service for IP networks," in *Proc. IEEE ICC '99*, vol. 1, 1999, pp. 671–675.
- [5] W.-S. Rhee, H.-S. Kim, K.-C. Park, K.-I. Kim, and S.-H. Kim, "Two phase edge-to-edge distributed measurement based admission control mechanism in large IP networks," in *Proc. IEEE GLOBECOM 2001*, vol. 4, 2001, pp. 2571–2575.
- [6] S. Krasser, H. Owen, J. Grimminger, H.-P. Huth, and J. Sokol, "Probing available bandwidth in radio access networks," submitted for publication.
- [7] F. Borgonovo, A. Capone, L. Fratta, and C. Petrioli, "VBR bandwidth guaranteed services over DiffServ networks," in *IEEE RTAS Workshop '99*, 1999.