Heterogeneous Real-time Traffic Admission Control in Differentiated Services Domains

S. Georgoulas[1], P. Trimintzios[1,2], G. Pavlou[1] and K. H. Ho[1]
[1]University of Surrey, UK, {S.Georgoulas, P.Trimintzios, G.Pavlou, K.Ho}@eim.surrey.ac.uk
[2]ICS-FORTH, Greece, ptrim@ics.forth.gr

***Abstract*- In Differentiated Services (Diffserv) domains, where services are provisioned on a per-class basis, admission control is an essential control factor in order to ensure that congestion is avoided and that the Quality of Service (QoS) requirements of individual flows are met. We consider traffic-engineered and provisioned IP Differentiated services domains able to support real-time traffic. We present a new Measurement-based Admission Control (MBAC) scheme that uses measurements of aggregate bandwidth only, without keeping the state of any per-flow information. In our scheme there is no assumption made on the nature of the traffic characteristics of the real-time sources, which can be of any heterogeneous nature. Through simulations we show that the admission control scheme is robust with respect to *traffic heterogeneity* and *measurement errors*. We also show that our approach compares favorably against other admission control schemes found in the literature.**

I. INTRODUCTION

QoS guarantees in a Diffserv domain can only be ensured through appropriate admission control policies [1] in order to control the amount of traffic injected into the domain. Admission control in such domains is a set of actions required at ingress nodes at the service instance establishment phase to check whether a service request is to be admitted or rejected. A new service instance should be admitted when the requested QoS can be satisfied without causing any QoS violation to the already established service instances. The various admission control approaches in the literature can be classified according to the method they use to decide if there are enough resources to accommodate the new service request. We can divide these methods into three categories: endpoint admission control (EAC), traffic descriptor-based admission control (TDAC), and measurement-based admission control (MBAC).

The first category of admission control is the endpoint admission control, which is based on some metric applied on probing packets sent by the end host/application along the transmission path before the flow is established [2]. The probing packets can be sent either at the same priority as flow packets (in-band probing) or at a lower priority (out-of-band probing). A requirement for the end-to-end route is to be the same for probing packets and flows. Setup delays may be high and, furthermore, simultaneous probing by many sources can lead to a situation known as *thrashing* [2]. That is, even though the number of admitted flows is small, the cumulative level of probing packets prevents further admissions, driving therefore the utilization to very low values. For in-band probing, *thrashing* additionally degrades the QoS perceived by the already established flows.

The second category is the traffic descriptor-based admission control, which is based on the assumption that a traffic descriptor, either deterministic or stochastic, is provided by the application for each flow requested prior to its establishment. This approach achieves high utilization when traffic descriptors used by the admission control scheme are tight [3]. Nevertheless, in practice, it suffers from several problems [1]. One of them is the inability of the application to come up with tight traffic descriptors before establishing the flow. This is especially true, when the bandwidth fluctuates over multiple time scales. Another problem is that this traffic descriptor and the associated QoS guarantee define a contract between the application and the domain. Therefore, the need to police based on this traffic specification arises, which is difficult for statistical traffic descriptors [1].

The last category is measurement-based admission control, which tries to avoid the problems of the other approaches by shifting the task of traffic characterization from the application to the network [4]. Instead of applications explicitly specifying their traffic descriptors, the network attempts to "learn" the characteristics of existing flows through real-time measurements. This approach has a number of advantages. First, the specified traffic descriptors can be very simple, e.g. peak rate, which can be easily policed. Second, an overly conservative specification does not result in over-allocation of resources for the entire duration of the service session. Third, when traffic from different flows is multiplexed, the QoS experienced depends often on their aggregate behavior, the statistics of which are easier to estimate than those of an individual flow. However, relying on measured only quantities for admission control raises a number of issues that need to be considered, such as the estimation errors, flow level dynamics and memory related issues [4].

In order for an MBAC scheme to be successful in practice, it has to fulfill several requirements [1, 3].

Robustness: An MBAC scheme must ensure that the requested QoS is provided. This is not trivial, since measurement inevitably has some uncertainty, potentially leading to admission errors. The QoS should also be robust to traffic heterogeneity, time-scale fluctuations (long-range dependency), as well as to heavy offered loads.

Resource utilization: The secondary goal for MBAC is to maximize resource utilization, subject to the QoS constraints for the admitted flows.

Implementation: The cost of deploying an MBAC scheme must be smaller than its benefits. In addition, the traffic characteristics required by the MBAC scheme should be easily obtained from the traffic sources and the network.

In this work we present a measurement-based admission control scheme for real-time traffic. We define as real-time traffic, sources that have a strict, usually small, *delay* and *jitter* requirement and a bounded, not necessarily too low, *packet loss rate* (PLR) requirement. In a Diffserv domain we assume that such real-time traffic is aggregated so that traffic from

sources composing each traffic aggregate will receive the same treatment over the entire domain. In Diffserv, core routers are stateless and agnostic to signaling. Per-flow state is only kept at ingress routers, while in the core network, traffic with similar QoS requirements is grouped in one of the engineered traffic classes and forwarded in an aggregate fashion.

We assume that the *delay* requirement of the traffic aggregate has been taken into account at the network provisioning stage. This means that the network provisioning processes configure appropriately small packet queues for the real-time traffic aggregate in order to keep the per-hop delay small. In addition, by controlling the routing processes to choose paths with constrained number of hops, we can keep the overall edge-to-edge delay under given bounds. Our assumption related to *packet loss*, is that packets are expected to be lost *only* at the first point of aggregation (network edge), which, according to [5], is currently considered as the most probable congestion point (bottleneck link) of a domain. We assume that further downstream inside the domain, real-time traffic aggregates are provisioned in a peak rate manner. This is feasible since, as stated in [6], in a common network configuration, backbone links are over-provisioned. According to [7], *jitter* can remain controlled in successive multiplexing queues as long as the flows are shaped to their nominal peak rate at the network ingress. Therefore, we assume that real-time traffic is conditioned and shaped based on the contracted peak rate. Furthermore, the deployment of non-work conserving scheduling in routers for the real-time traffic class can be beneficial for controlling *jitter* [8].

We also assume that the interior of the Diffserv domain has been provisioned and engineered in this way in order to support the real-time traffic aggregates. As a result of the provisioning process, and taking into account the routing behavior, at each ingress node, we can have an estimate of the minimum bandwidth available for the real-time traffic aggregate from that ingress to each of the corresponding egress nodes. This available bandwidth is the *basis* for our admission control scheme, which is employed at the edge (ingress) node of the first Diffserv aggregation point. Our assumptions imply that our admission control scheme does not induce any states in the core network, which is desired for scalability and resilience reasons, and it is also proven to be a resource-efficient approach if resilience against network failures is required [9].

The rest of this paper is organized as follows. Section 2 presents our measurement-based admission control scheme. In Section 3 we evaluate the performance of our scheme comparing it to other approaches found in the literature. Finally, in Section 4 we conclude, summarizing our findings.

II. Admission Control Scheme

In this section we will present our Measurement-based Admission Control scheme, applicable to real-time sources that are able to provide *only* a single traffic descriptor, their peak rate. Given the diversity of Internet-based applications that have real-time requirements, the use of more complex traffic descriptors in admission control, as stated in [10], to accurately characterize source traffic, is neither necessary nor plausible. Therefore, we assume that the only available traffic descriptor to use is the source's peak rate. This traffic descriptor is easy to police and even if not available, for sources described by a token bucket filter (r,b) an estimate $\hat{p}$ of it can be derived [10] using the equation:

$$\hat{p} = r + b/U \quad (1)$$

where U is a user-defined averaging period.

Our scheme uses the bufferless statistical multiplexing approach. Bufferless multiplexing is very attractive for real-time traffic since it ensures that the traffic experiences minimal delay. In addition, the dynamics leading to an overload event in a bufferless system are much simpler than those of a buffered system [11]. The main disadvantage of using a buffer is that overflow probability depends significantly on flow characteristics [12] and can only be tightly controlled if these characteristics are known. Moreover, in this case, provisioning needs to account for statistical variations in the traffic mix as flows arrive and terminate. On the other hand, buffered multiplexing allows higher utilization for the same loss rate [12] but requires more complex traffic management and is not as robust with respect to flow characteristics as bufferless multiplexing. We need to stress that bufferless multiplexing is just an abstraction [12]. For packetized traffic, as in IP networks, a small buffer for packet scale queuing is needed to account for coincident packet arrivals from distinct flows [7].

According to [13], when the effect of statistical multiplexing is significant, the distribution of the stationary bit rate can be accurately approximated by a Gaussian distribution. In [14] it is suggested that the aggregation of even a fairly small number of traffic streams is usually sufficient for the Gaussian characterization of the input process. In that case, the effective bandwidth of the multiplexed sources is given by:

$$C \simeq m + a'\sigma \text{ with } a' = \sqrt{-2\ln(\varepsilon) - \ln(2\pi)} \quad (2)$$

where m is the mean aggregate bit rate, σ is the standard deviation of the aggregate bit rate and ε is the upper bound on allowed loss probability.

A. Algorithm for Admission Control

In a Diffserv domain we assume that the real-time traffic aggregate is provisioned and engineered in such a way that at minimum C_{total} bandwidth is available edge-to-edge. Every time a source wants to establish a service instance, it signals this to the ingress node through some resource reservation protocol. A similar assumption can be made for the service termination. If the latter is not explicitly signaled, an alternative option could be to use a time-out period as an indication of the service termination. In any case, at each point in time, the MBAC process at each ingress point knows the number of active sources.

When a new service request arrives, we need to decide whether or not to allow the source to send traffic using the real-time traffic aggregate resources until the known egress point. Initially, we need to calculate an appropriate time period, the *measurement window*, within which we need to take and use measurements for bandwidth usage estimations. The measured parameters are the mean rate of the offered load,

$M_{measured}$, and the variance of the offered load, $\sigma^2_{measured}$, at the output queue of the ingress node. Having the measurements and the peak rate p_{new} of the new source, and by making the worst case assumption that the new source will be transmitting at its peak rate, we compute the estimated bandwidth C_{est} as follows:

$$C_{est} = M_{measured} + p_{new} + a'_{PLR}\sqrt{\sigma^2_{measured}} \quad (3)$$

where a'_{PLR} is computed as in (2), based on the target PLR bound of the real-time traffic aggregate. This value C_{est} is the estimated bandwidth used in the admission control criterion.

B. Measurement Window Estimation

We define the measurement window w, as the time interval within which the offered load is taken into account for deriving the required measurements. In a similar fashion to [15], we use the following expression for the measurement window:

$$w = \max(DTS, w') \quad (4)$$

In (4), DTS represents the Dominant Time Scale. DTS is the most probable time scale over which overflow occurs. In [14], the authors describe a systematic way to derive DTS using real-time measurements, with the assumption that the input process to the multiplexing point in the network is Gaussian. This is by definition our assumption when employing (2), therefore we use this method in order to estimate the DTS. DTS, as computed in [14], is a function of the mean rate, the variance of the offered load and the output buffer size. The reader should recall that even though we employ the bufferless multiplexing approach, a small output buffer is still required for packet scale queuing, as explained in the previous section. This value for the output buffer is involved in the estimation of the DTS.

Let w' represent the mean inter-departure delay [4], defined as follows (Little's formula):

$$w' = \frac{h_{avg}}{N_{active}} \quad (5)$$

where N_{active} is the number of simultaneously active sources and h_{avg} is their average duration.

Since we assume that the service establishment and termination is signaled to the ingress nodes, the average duration of the sources can be easily obtained and updated.

That is we select as measurement window the mean inter-departure delay, i.e., the time interval within which the system can be considered stationary -no flow departures-, unless this time interval is not long enough to capture the time-scale fluctuations of the aggregate traffic stream. This can happen in case of long-range dependent traffic. In this case and in order to enable the network to react to these traffic fluctuations, we use DTS as the value of the measurement window.

C. The Admission Control Criterion

Given that the allocated bandwidth for the real-time traffic aggregate from edge-to-edge is C_{total}, and having computed the estimated bandwidth C_{est}, the admission control criterion in our scheme becomes:

$$\begin{aligned} &If \quad (C_{est} \times APF) \le C_{total,} \quad admit \\ &If \quad (C_{est} \times APF) > C_{total,} \quad reject \end{aligned} \quad (6)$$

where *APF* is an Admission Policy Factor we involve in the admission control criterion. The use of *APF* reflects the provider's policy on how strict the admission control should be. The decisions for setting the *APF* can be based on simple heuristics or ad hoc engineering methods. In the following section we describe an example approach for setting *APF*, in which we take into account two issues that challenge the effectiveness of any MBAC scheme:
(a) the traffic source heterogeneity, and
(b) the effect of measurement errors.

D. A Heuristic for Setting Admission Control Policies

The reason for introducing *APF* is to reflect the provider's policies. This means that appropriately tuning the *APF* can lead to a more conservative or a more relaxed admission control criterion. In our case we give a heuristic formula for *APF* with which we address two important issues that need to be taken into account in the admission control decision.

The first issue is that the aggregate traffic stream might have characteristics that do not suit the effective bandwidth formula (2). This, for instance, can happen if the stream is composed of a small number of very bursty connections with high peak rates and low utilizations [13].

To account for this, we use an exponential ON/OFF source, with mean and standard deviation (m_{ref}, σ_{ref}) as a model source for engineering reasons (*reference source*). The reason for the specific selection is that exponential ON/OFF sources are representative models for VoIP traffic, which is likely to be a big part of the traffic carried by real-time traffic aggregates and their traffic characteristics suit the effective bandwidth formula (2). Furthermore, exponential ON/OFF sources are short-range dependent, which means that their traffic characteristics are more easily captured within the given measurement window. We define as *reference trunks* (T_{ref}) the number of simultaneously established reference sources that can fit in C_{total}, according to (2), for a given bound on packet loss rate.

When a new request arrives, having measured the mean rate $M_{measured}$ and the variance $\sigma^2_{measured}$ of the offered load, we calculate the number N_m of the reference sources, whose aggregate mean rate is equal to or greater than $M_{measured}$. We also calculate the number N_σ of the reference sources, whose aggregate variance is equal to or greater than $\sigma^2_{measured}$. That is, N_m and N_σ satisfy the following relationships:

$$N_m = \left\lceil \frac{M_{measured}}{m_{ref}} \right\rceil \text{ and } N_\sigma = \left\lceil \frac{\sigma^2_{measured}}{\sigma^2_{ref}} \right\rceil \quad (7)$$

Having estimated N_m and N_σ, we compute their mean value N_{ref}:

$$N_{ref} = (N_m + N_\sigma)/2 \quad (8)$$

This value represents a rough estimate of the number of reference sources that produce, within the measurement

window, load with characteristics (mean rate and variance) similar to the ones measured. To compensate for the above, we set *APF* to be proportional to the quantity (N_{ref} / T_{ref}).

The second issue that needs to be taken into account with the policy factor is the effect of measurement errors. As shown in [4], the *certainty equivalence* assumption, i.e., that the measured parameters represent the real traffic, can heavily compromise the performance of an MBAC scheme. The stringent the PLR requirement, the easier it is to violate it due to measurement errors. In the case where only aggregate bandwidth information is available through measurements, as in our scheme, the degradation in performance can be mainly attributed to errors in the estimation of the variance [1]. With non-negligible probability the variance can be significantly underestimated. To compensate for the measurement uncertainty, we proceed as follows: given (2), for a specific target PLR, we set APF to be proportional to the quantity $\frac{\sqrt{-2\ln(PLR)-\ln(2\pi)}}{\sqrt{-2\ln(PLR_{ref})-\ln(2\pi)}}$.

That is, we inflate the part of equation (2) that relates to the variance estimation, based on a reference PLR level. By setting PLR_{ref} to be larger than PLR, we ensure that the more stringent the PLR requirement, the greater the value of this quantity. This reference PLR can be set by policy to adjust the conservativeness of the MBAC scheme.

Combining the two aforementioned quantities, the final expression for the admission policy factor that can be used is:

$$APF = (N_{ref} / T_{ref}) * \frac{\sqrt{-2\ln(PLR)-\ln(2\pi)}}{\sqrt{-2\ln(PLR_{ref})-\ln(2\pi)}} \quad (9)$$

We set $APF = 1$ whenever the equation above results to be less than 1. That means that we use APF in a conservative way in the admission control criterion. The admission policy factor can be considered as a tuning parameter. Even though we derive *APF* somehow heuristically, based on intuition rather than strict mathematical analysis, one should take into account that all MBACs employ additional admission policy tuning parameters [1, 16] because it is not possible to completely decouple performance from traffic characteristics.

III. Performance Evaluation

In order to evaluate the performance of our admission control scheme, we run simulations using the network simulator *ns-2* [17], with the dumbbell topology of Fig. 1.

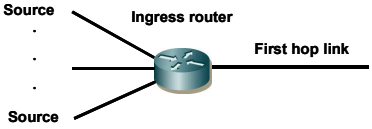

Fig. 1. Simulation topology.

We assume that the sources connect to the ingress node through links with negligible congestion. Even if the sources originate from local area networks (LANs), this assumption can still be considered valid, since LANs have large bandwidth enough to handle a number of sources, larger than that of wide area networks in most real network situations [18].

We set the reference PLR equal to 0.01 and we configure the output queue for the real-time traffic class to hold a maximum of 5 packets and to be served in a non-work conserving scheduling manner. As a *reference source*, we use an exponential ON/OFF source with a peak rate of 64kbps and mean durations for the ON and OFF periods 1.004sec and 1.587sec respectively [19]. We use scenarios with the target bound on packet loss rate for the aggregate real-time traffic equal to 0.01 and 0.001.These bounds represent typically acceptable PLR values for the VoIP service and for real-time applications in general, according to [20, 21]. We need to stress here that if the real-time applications require different PLR (or QoS in general) targets to be met, then multiplexing them in a single class would require for the most stringent QoS requirement among them to be met. That would lead to severe underutilization of resources. In such a case, applications with different QoS requirements should be multiplexed in different classes based on the value of the requested QoS. For our simulations we assume that the real-time applications have the same PLR requirement and can, therefore, be multiplexed in a single class.

We set the output link capacity allocated to real-time traffic to correspond to T_{ref} equal to 100. This means that the output link capacity is set equal to 3.33Mbps for the target PLR 0.01 case and equal to 3.56Mbps for the target PLR 0.001 case. In a real network situation, unused capacity of the real-time traffic class would be fully available to a lower priority, e.g. best effort, traffic, so there would be no waste incurred by this partitioning. All the results given in this section are based on averages of simulations for 20 randomly chosen seeds, each for a total of 4100 seconds, using the first 500 seconds as a warming-up period.

In order to test the *robustness* of the scheme with respect to traffic *heterogeneity* and *long-range dependency*, we use both VoIP and Videoconference traffic sources. For VoIP traffic we use an ON/OFF source model with exponentially distributed ON and OFF times, having a peak rate of 64kbps. The mean durations for the ON and OFF periods are 0.350sec and 0.650sec respectively [22]. The active time of the VoIP sources is exponentially distributed with an average of 300sec. For Videoconference traffic we use an H.263 coded trace from [23] with average rate 64kbps and peak rate 332.8kbps. The H.263 format has been widely employed to model videoconference traffic, e.g., see [24]. The active time of the Videoconference sources is exponentially distributed with an average of 180sec. For both VoIP and Videoconference sources, the activation processes are Poisson arrival processes. For the cases where both VoIP and Videoconference sources are employed (mixed traffic), the averages of their activation rates follow a ratio of 2:1.

In order to test the *robustness* of the scheme with respect to *offered load*, we test varying load conditions ranging from 0.5 to 5, where the value 1 (*reference load)* corresponds to the average load that would be incurred by a source activation rate equal to 1000 VoIP sources/hour.

In order to compare the performance of our scheme, which we call MBAC-GEO, against other existing proposals, we

implement three other algorithms. The first algorithm is an MBAC scheme described by Zukerman et al in [25] as Rate Envelope Multiplexing (REM), with adaptive weight factor and no histogram update. The reasons for the selection of the specific MBAC scheme (we call it MBAC-ZUK) for comparison with our scheme are that: (a) REM also makes the zero buffer approximation with respect to statistical multiplexing and (b) implementation-wise, in a similar fashion to our scheme, it requires only aggregate bandwidth measurements and the peak rate of the sources requesting admission in order to derive the admission control decision.

The second algorithm is an EAC scheme described by Karlsson et al in [26]. In order to test this scheme (we call it EAC-KAR) we implement an additional lower priority queue for the probing packets (out-of-band probing) that can store, as proposed in [26], a single probe packet and which is only served when the higher priority real-time traffic queue is empty. As in [26], we set the probing rate equal to the peak rate of the source requesting admission, we consider probe durations of 0.5sec up to 5sec, and we also assume that there is no latency involved between the probing phase completion and the admission control decision.

The third algorithm is a simple TDAC Peak Rate Allocation scheme (we call it TDAC-PRA) that only admits a new source if the following condition is satisfied:

$$\sum p_i + p_{new} \leq C_{total} \quad (10)$$

where $\sum p_i$ is the sum of peak rates of the already established sources. With this scheme, there are no losses, since it does not account for any statistical multiplexing.

As stated in [27], any admission control scheme must address the trade-off between *packet loss* and *utilization*. Therefore for performance evaluation we use these two metrics, together with the average admission rejection rate.

In our simulations we consider two cases for the mixture of traffic sources that request admission: (a) Videoconference sources only and (b) Mixed VoIP and Videoconference sources.

For TDAC-PRA we do not show PLR results because it is constantly zero. For EAC-KAR, the results shown are for a probe duration of 2 seconds, which gives the best trade-off between packet loss and utilization. For lower probing durations we observe violation of the target PLR, whereas for higher probe durations we enter the *thrashing* region.

A. Videoconference Sources

The performance results for videoconference traffic sources are shown in Fig. 2 and 3.

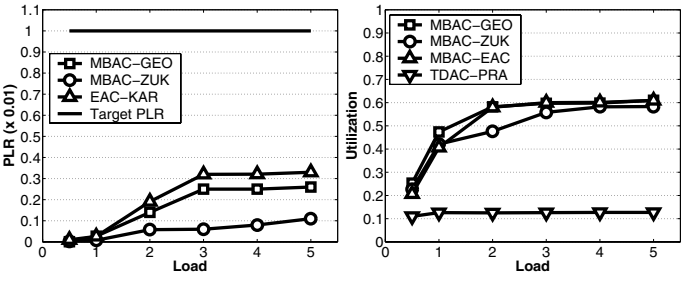

Fig. 2. Achieved PLR and utilization for target PLR 0.01.

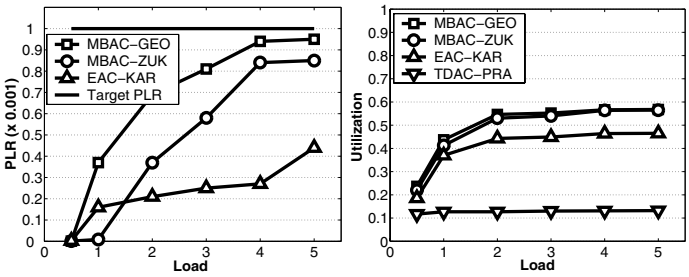

Fig. 3. Achieved PLR and utilization for target PLR 0.001.

For videoconference traffic, all schemes achieve the target PLR for all loading conditions. For target PLR 0.01, all schemes are unnecessarily conservative, which can be partly attributed to the stringent admission control criterion (all schemes make the worst case assumption that the new source will be transmitting at its peak rate) and the high peak rate of the videoconference sources compared to their average rate. Regarding utilization, the performance of MBAC-GEO is, on average, better than that of EAC-KAR and slightly better than that of MBAC-ZUK.

The reader should recall at this point that the objective is not to achieve the lowest PLR possible, but to keep the achieved PLR within the limits of the target PLR, while maximizing utilization. For example, for TDAC-PRA the achieved PLR is zero, but the utilization is significantly lower than any of the other three algorithms, and because of the bursty nature of the H.263 videoconference traffic it does not exceed 15%.

B. Mixed VoIP and Videoconference Sources

The performance results for mixed heterogeneous traffic sources are shown in Fig. 4 and 5.

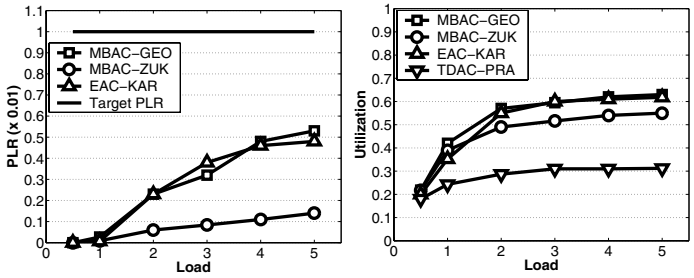

Fig. 4. Achieved PLR and utilization for target PLR 0.01.

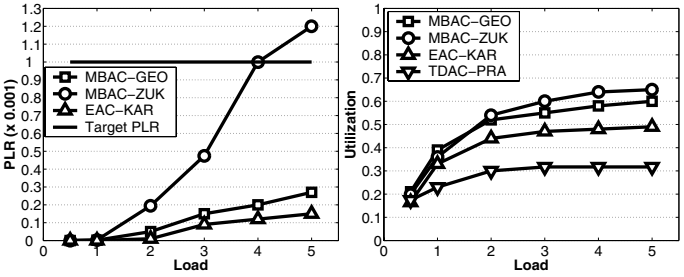

Fig. 5. Achieved PLR and utilization for target PLR 0.001.

For mixed traffic, all schemes achieve the target PLR 0.01. MBAC-ZUK is more conservative than MBAC-GEO and EAC-KAR, achieving therefore lower utilization. For target PLR 0.001, MBAC-GEO and EAC-KAR achieve this PLR for all loading conditions with MBAC-GEO being less

conservative, achieving a higher utilization. MBAC-ZUK violates this PLR for loading conditions more than 4 times the *reference load*, even though the no histogram update method used in our implementation of MBAC-ZUK is the most conservative approach among all the other variations [25]. This means that the tuning parameters involved in MBAC-ZUK should be reconfigured in a trial and error fashion in order to achieve the target PLR for all loading conditions. We need to stress here that EAC-KAR also employs a tuning parameter, which is the probe duration, and which we have to vary from 0.5sec up to 5sec in order to find its optimal value (2sec) for the simulated cases and loading conditions.

TDAC-PRA achieves much higher utilization compared to the previous case because of the existence of the less bursty VoIP sources in the traffic mix, but still significantly lower than any of the other three schemes.

The averages of utilization and admission rejection rate for all simulated cases and loading conditions are shown in Fig. 6.

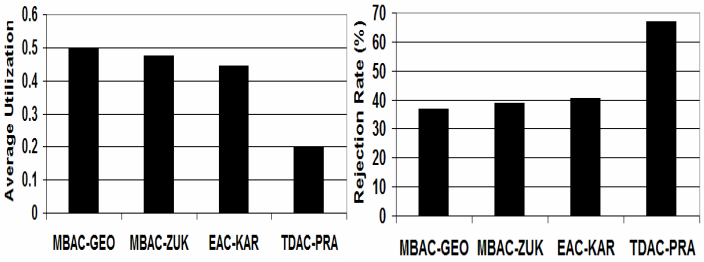

Fig. 6. Average utilization and admission rejection rate.

MBAC-GEO achieves, on average, the highest utilization and the lowest admission rejection rate among all schemes, while satisfying the target PLR.

In all cases examined, for both MBAC-GEO and MBAC-ZUK we observe an increase in the achieved PLR for higher load conditions. This is anticipated [4] because they both rely on measurements, so every new admission request has the potential of being a wrong decision. This means that a high source activation rate is expected to have a negative effect on performance. The same holds for EAC-KAR since we have not entered the *thrashing* region.

IV. Conclusions

In this paper we propose a measurement-based admission control scheme for heterogeneous real-time traffic in Diffserv domains. We assume that an instance of our MBAC scheme runs at every ingress node. We show through simulations that the scheme is *robust* to *traffic heterogeneity*, time-scale fluctuations and heavy offered loads. The scheme can satisfy the packet loss rate requirement in all cases despite the effect of *measurement errors* and without requiring any reconfiguration of its parameters. Furthermore, the scheme achieves satisfactory *utilization* and compares well against existing admission control approaches for the same simulation setup. We should also mention that our scheme is easy to *implement* since it does not require any per-flow information state. The required traffic characteristics are the peak rate of the traffic source requesting admission and the mean rate and the variance of the aggregate real-time traffic load at the output queue of the ingress node where the MBAC instance runs.

References

[1] M. Grossglauser and D. Tse "A Time-Scale Decomposition Approach to Measurement-Based Admission Control", IEEE/ACM Transactions on Networking, August 2003.

[2] L. Breslau, E. Knightly, S. Shenker, I. Stoica and Z. Zhang "Endpoint Admission Control: Architectural Issues and Performance", SIGCOMM 2000, Stockholm 2000.

[3] G. Mao and D. Habibi "Loss Performance Analysis for Heterogeneous ON-OFF Sources with Application to Connection Admission Control", IEEE/ACM Transactions on Networking, February 2002.

[4] M. Grossglauser and D. Tse "A Framework for Robust Measurement-Based Admission Control", IEEE/ACM Transactions on Networking, June 1999.

[5] V. N. Padmanabhan, L. Qiu and H. J. Wang "Server-based inference of Internet Link Lossiness", IEEE INFOCOM 2003.

[6] G. Iannaccone, M. May, and C. Diot "Aggregate Traffic Performance with Active Queue Management and Drop from Tail", Computer Communications Review, July 2001.

[7] T. Bonald, A. Proutiere and J. Roberts "Statistical Performance Guarantees for Streaming Flows using Expedited Forwarding", IEEE INFOCOM 2001.

[8] M. Mowbray, G. Karlsson ant T. Kohler "Capacity Reservation for Multimedia Traffics", Distr. Syst. Eng., 1998.

[9] M. Menth "Efficient Admission Control and Routing for Resilient Communication Networks", PhD Thesis, Univ. of Wurzburg, July 2004.

[10] S. Floyd "Comments on Measurement-based Admission Control for Controlled-Load Services", July 1996, Lawrence Berkeley Laboratory Technical Report.

[11] D. Tse and M. Grossglauser "Measurement-based Call Admission Control: Analysis and Simulation", IEEE INFOCOM 1997.

[12] T. Bonald, S. Oueslati-Boulahia and J. Roberts "IP traffic and QoS control: the need for a flow-aware architecture", World Telecommunications Congress, September 2002.

[13] R. Guerin, H. Ahmadi, and M. Naghshieh "Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks", IEEE Journal on Selected Areas in Communications, September 1991.

[14] D. Eun and N. Shroff "A Measurement-Analytic Approach for QoS Estimation in a Network Based on the Dominant Time Scale", IEEE/ACM Transactions on Networking, April 2003.

[15] S. Belenki "An Enforced Inter-Admission Delay Performance-Driven Connection Admission Control Algorithm", ACM SIGCOMM 2002.

[16] L. Breslau, S. Jamin and S. Shenker "Comments on the Performance of Measurement-Based Admission Control Algorithms", IEEE INFOCOM 2000.

[17] K. Fall and K. Varadhan "The *ns* manual" (www.isi.edu/nsnam/ns/ns_doc.pdf).

[18] A. Bilhaj and K. Mase "Endpoint Admission Control Enhanced Systems for VoIP Networks", IEEE SAINT 2004.

[19] C. Chuah, L. Subramarian, and R. Katz "Furies: A Scalable Framework for Traffic Policing and Admission Control", May 2001, U.C Berkeley Technical Report No. UCB/CSD-01-1144.

[20] I. Frigui "Services and Performance Requirements for Broadband Fixed Wireless Access", IEEE P802.16 Broadband Wireless Access Working Group.

[21] T. Chaded, "IP QoS Parameters", TF-NGN, November 2000.

[22] I. Habib and T. Saadawi "Multimedia Traffic Characteristics in Broadband Networks", IEEE Communications Magazine, July 1992.

[23] http://www-tkn.ee.tu-berlin.de/research/trace/trace.html

[24] S. Atallah, O. Layaida, N. Palma and D. Hagimont "Dynamic Configuration of Multimedia Applications", IFIP/IEEE MMNS 2003.

[25] T. Lee and M. Zukerman "Admission Control for Bursty Multimedia Traffic", IEEE INFOCOM 2001.

[26] V. Elek, G. Karlsson and R. Ronngren "Admission Control based on End-to End Measurements", IEEE INFOCOM 2000.

[27] R. Gibbens and F. Kelly "Measurement-based connection admission control", 15th International Teletraffic Congress, June 1997.