Measurement-based admission control: algorithms, evaluation, and research perspectives

Hans van den Berg and Michel Mandjes

Abstract—Currently, in the context of ATM and IP networks there is much interest in measurement-based admission control (MBAC), i.e., admission control based on real-time traffic measurements. We present an overview of the literature on MBAC, aim to categorize these references, and identify their main characteristics. Two MBAC algorithms proposed in literature are selected, and by means of simulation we verify their implementability and study their performance under various traffic scenarios (both for artificially generated traffic and for traces of real traffic). We provide reflections on the use of MBAC, and identify issues that require additional research.

Key words: Measurement Based Admission Control (MBAC) — Quality of Service — ATM and IP networks — simulation

I. INTRODUCTION

A major challenge when operating an integrated services broadband network is the control of the balance between 'quality' and 'efficiency'. The users of the network should be delivered a sufficient Quality of Service (QoS), while economic viability demands that the network operates at a fairly efficient level (in terms of utilization of resources). In order to achieve this, for emerging network technologies such as ATM (asynchronous transfer mode) and IP (Internet protocol), a set of traffic control mechanisms has been developed.

One of the most important traffic control mechanisms is admission control (AC). A candidate network user submits a request, consisting of an upper bound on the traffic he will send (commonly called the 'traffic contract') and a specification of the required QoS (generally in terms of the maximum permitted transfer delay and fraction of data lost). The network decides by means of the AC algorithm whether or not the request can be accepted. In other words, it is checked whether sufficient resources are available in the network to meet the QoS requirements of the new request and of the requests already admitted. Once the request is admitted, the traffic stream's compliance with the traffic contract is enforced by the policing function.

The traffic contract is usually only a rough 'envelope' of the traffic actually sent by the user. This is mainly due to the fact that the traffic contract typically consists of only three parameters (peak rate, mean rate and burst size, see e.g. [15], [23]), which is simply too few to give an accurate description of the actual traffic pattern. In the second place, it is in general hard for a user to a priori describe his traffic pattern, as it very much depends on the specific actions the user undertakes during his session. Hence, a reliable AC based on the traffic contract — usually called 'static' AC — will be very conservative, in that it leads to a low utilization of network resources.

Based on the above line of reasoning, clearly static AC should only be used in the case that very strict QoS guarantees are required. Fortunately, a significant set of services can cope with temporary QoS degradations (e.g., data services using TCP/IP) and can be subjected to a less conservative admission control approach. In this way the gap between the traffic contract and the real traffic profile can be exploited: one can allow more flows to the network than justified by AC based on worst-case profiles.

These so-called 'overbooking' [4] schemes are already commonly used in Frame Relay (data) networks. It should be emphasized that overbooking has a couple of important drawbacks. In the first place the 'overbooking factor' is not universal, as it is clearly user-specific and application-specific. Secondly, overbooking is basically extrapolating historic measurements, which obviously does not exclude possible service degradation.

The performance of such an overbooking scheme can be improved by making regular updates of the 'overbooking factor'. An extreme variant is to base the AC on real-time traffic measurements. This novel approach to AC is usually called measurement-based admission control (MBAC; also: 'dynamic AC'), see e.g. [13], [14], [16]. In its ideal form, (i) the MBAC's operations are kept quite simple, (ii) the efficiency is considerably increased (compared to static AC), and (iii) the QoS is controlled relatively well. The last
few years MBAC gets much attention in teletraffic literature, and several algorithms have been proposed. Some of these algorithms show promising results, but they still have their limitations and are not practically implementable yet.

We adhere to the point of view that one should use both a static AC (for traffic with strict QoS requirements) and a measurement based AC (for traffic with 'loose' QoS requirements) in an integrated services network. An effective combination of these two AC concepts is enabled by appropriate packet scheduling mechanisms like priority queueing and Weighted Fair Queueing (WFQ).

Contribution. Research in the area of MBAC has just taken off; we present a brief overview of the most relevant references. We aim to categorize them, and identify their main characteristics.

Then we give a numerical study, which serves two major goals: (i) we verify (by simulation) whether it is practically feasible to implement the MBACs proposed in the literature, and (ii) investigate the efficiency gain that may be obtained in practical situations. For that reason we have selected two MBAC schemes that have attracted attention [2], [16] and assess the performance of these algorithms under various traffic conditions. In particular, we use both 'artificial' input (simulated sources, like on-off traffic) and traces of real traffic. We observe a significant efficiency gain with respect to static AC. Remarkably, we find that both algorithms perform better for heavy tailed traffic than for exponential traffic - notice that these heavy tails cause traffic to exhibit long range dependence [18].

Finally, we identify a number of (practical and theoretical) issues that have to be settled in order to enable MBAC in a real-life network.

The paper is organized as follows. Section II presents an overview of the literature on MBACs. Section III describes two algorithms; we have implemented them in a simulation program. Section IV deals with a discussion on the simulation output and the comparison with a static AC. This section also includes a number of theoretical considerations based on an analytical model. Section V gives conclusions and directions for future research.

II. OVERVIEW OF LITERATURE

During the past years a significant number of papers on MBAC has appeared. Below we review the most significant among them. Two notions play a crucial role.

- Jamin et al. [16] introduced the separation between the admission criterion and the measurement procedure. The admission criterion determines on the basis of a number of traffic characteristics (of the existing flows and the new flow) whether or not to accept the new flow. The measurement procedure captures the required traffic characteristic from the flows that are currently present (possibly in combination with the a priori traffic descriptor).
- Grossglauser and Tse [13] introduce the concept of a certainty equivalent (CE) method. These methods use a static AC algorithm, but insert measured quantities rather than a priori known traffic descriptors (and these measured quantities are assumed to be the 'real ones').

Certainty equivalent methods. The attractive feature of CE methods is that one can reuse existing static ACs. However, as pointed out nicely in for instance [13], CE methods will give too optimistic results. This is due to the fact that the measured quantities themselves are random variables and therefore incur additional uncertainty. Inserting these quantities into the static AC as if they were the 'real ones' might therefore lead to QoS degradation.

Quite a number of MBACs that appeared in the literature are in fact CE methods. Mostly the above mentioned aggressive properties are compensated in an indirect way by a conservative measurement procedure. An example of such a method is Jamin et al. [16]. In this algorithm measured rates (instead of the leaky bucket parameters, that are upper bounds of the rates) are inserted into the 'Parekh and Gallager deviation bounds' [20]. A new connection is admitted as long as the delay estimate (obtained in this way) does not exceed the required value. The aggressive properties of this algorithm are compensated by measuring the rate in a conservative manner, see Section III.A. Brichet and Simonian [2] follow a similar procedure: they insert measured rates into an effective bandwidth formula and use the exponentially weighted moving average (EWMA) measurement procedure. Reisslein's study [21] is comparable to [2]: the admission policy works in a CE manner, and the measurement scheme is based on EWMA.

As said above, the measurement procedures should be tuned such that conservativeness is recovered. These procedures are parametrized by a couple of
tuning knobs, like window sizes and the exponential weight of historic data. A successful implementation of the 'CE approach' (a combination of a static AC and a conservative measurement procedure) requires a mechanism for choosing suitable values for these tuning parameters. That seems to be far from trivial, in particular due to the fact that the proper choice depends on the traffic characteristics. Notice that this was just what MBAC should avoid! However, if the choice of the tuning parameters is only modestly affected by the traffic pattern, the approach is still viable. We will extensively comment on this issue in Section IV.A. Casetti et al. [3] advocate an algorithm that adjusts the tuning parameters on the basis of the traffic offered.

An interesting conjecture was postulated by Jamin and Shenker [17]. They consider the flow blocking/packet loss curve — obviously if the packet level criterion is stringent (low packet loss), substantial blocking will occur; similarly high packet loss will go with low blocking. In [17] it is stated that every well-tuned MBAC has (for given input traffic) the same flow blocking/packet loss curve. Notice that this does not mean that all MBACs are equivalent, as there are more performance measures than just packet loss and blocking (delay for instance), and apart from that the MBACs may differ in measurement effort required.

Also in Duffield et al. [6], [7] a CE procedure is presented: it is tried to capture the aggregate traffic's cumulant function, enabling to calculate the asymptotics of the loss ratio. They however do not compensate by a conservative measurement procedure, and therefore the approach seems to be quite 'dangerous'. A similar remark holds for the method of Courcoubetis et al. [5]. They measure the buffer occupancy, and extrapolate the resulting empirical distribution function. Apart from that, an implicit assumption is exponential decay of the buffer content distribution, which does not hold with long-range dependent input. Gibbens and Kelly [11] provide a family of admission criteria (all of them based on Chernoff bound arguments), but do not aim to provide measurement procedures.

Methods that take into account the randomness of the measurements. The papers of Grossglauser and Tse [13], [14], Duffield [9] and Gibbens, Kelly, and Key [12] are essentially different from the above in that they depart from using CE methods. They recognize that overload is due to jointly occurring 'misleading measurements' (giving an overly optimistic impression of the momentary load) and rare behavior in the period after the measurement. CE methods neglect the first type of error.

Under weak assumptions on the asymptotic regime (heavy load, a single connection's bandwidth being small compared to the link rate) [13] characterizes the error made by using a CE method. The same is done in [9] for the regime where loss is rare and large deviations theorems are applicable; it is shown that the presence of correlation between the samples may degrade the performance of the CE-based MBAC even more.

As said above, the MBACs of [13], [14], [9], and [12] explicitly take into account the stochastic character of the measurement itself. In one of the algorithms of [12] this stochastic nature is dealt with in a Bayesian way. The same holds for [9]; the loss probability is written as the sum of the loss probabilities for a specific probability model, weighted by the probability that that specific model was the real one (given the measurements). The interesting feature of [14] is that both burst-scale fluctuations as well as call-scale fluctuations play a role in the algorithm: for burst-scale fluctuations bandwidth has to be reserved on top of the connections' mean rates, but the slow call-level fluctuations can be exploited effectively by the MBAC.

III. DESCRIPTION OF SELECTED MBAC ALGORITHMS

This subsection focuses on the algorithms that we have implemented in a simulation program, viz. the algorithm of Jamin et al. [16] and an effective bandwidth (EB) based algorithm, which can be considered as a variant of [2]. Before doing that, we first sketch the situation considered.

We assume that the network supports two traffic flow types having different QoS requirements: 'predictive' traffic and 'guaranteed' traffic. Guaranteed traffic has stringent delay (and loss) requirements. Predictive traffic has 'soft' QoS requirements mainly based on a delay bound $D_P$ (i.e., it is required that typically more than, say, 99.9% of the packets has a delay smaller than a relatively large, predefined delay bound). The service differentiation between guaranteed and predictive traffic is enabled by the use of priority buffers at the switch/router output ports. A small buffer is used for the guaranteed traffic; it is served with high priority. A second, much larger buffer, is used to store the packets of the predictive
traffic flows; this buffer is served if the high priority buffer is empty, see Figure 1.

![Diagram of Priority Buffer Architecture](attachment://image.png)

**Figure 1: Priority buffer architecture for integration of guaranteed and predictive traffic**

In fact, both MBACs considered are based on the idea to use a conservative (not very efficient, static) AC for the guaranteed traffic flows and to use a more efficient (measurement-based) AC for the predictive traffic flows, for which the QoS requirements are not that hard. In this way we can exploit the bandwidth allocated to guaranteed flows, but not actually used by them: we can allow additional predictive traffic flows. Note that (due to the priority queuing at the switch/router output ports) the allowance of (too many) additional predictive flows will never harm the QoS received by the guaranteed traffic flows. Consequently this approach enables an efficient (overall) use of link bandwidth, while still guaranteeing the stringent QoS requirements of the guaranteed traffic.

It is assumed that the traffic flows are regulated by a token bucket filter, which determines the traffic contract parameters peak rate \( p \), maximum burst size \( b \) and ‘long term maximum average rate’ (sustained rate) \( r \). Note, that these are the parameters that are used for the characterization of e.g. SBR and RSVP connections in ATM and IP, respectively [15], [23]. In the next two subsections we will introduce in more detail the algorithms that we have implemented.

### A. Algorithm by Jamin et al.

**The admission rules.** On-line measurements are used to get estimates for the actual aggregate rate \( \nu_G^m \) of the guaranteed traffic and the actual aggregate rate \( \nu_P^m \) of the predictive traffic. Now, if a new guaranteed flow (with sustained rate \( r \)) arrives, the algorithm first checks whether there is enough bandwidth available:

\[
\nu_G^m + r < uC \quad \text{and} \quad \nu_G^m + \nu_P^m + r < C.
\]

(\( \nu_G^m \) denotes the sum of the declared sustained rates of the guaranteed traffic flows already in progress; \( u \) is the link utilisation target – typically \( u = 90\% \) – a kind of ‘safety margin’). Next, assuming that an amount of bandwidth \( C - \nu_G^m - r \) is available for predictive traffic and using measurements of the delays currently observed by the predictive packets, the algorithm checks - based on the ‘Parekh and Gallager bound’ [20] – whether the delay requirement of the predictive traffic will be met. If all these conditions are satisfied, the new flow is accepted. In the case of the arrival of a new predictive flow, the first condition, i.e., \( \nu_G^m + r < uC \), can be neglected (predictive flows have no impact on QoS of guaranteed flows); in the delay condition for predictive traffic, the burst size \( b \) of the new predictive flow is taken into account.

**The measurement process.** The measurement process can be roughly described as follows. Consecutive measurement windows of length \( T \) are used to get estimates for the packet delay and for the aggregate rates \( \nu_G^m \) and \( \nu_P^m \). During a measurement window, the predictive and guaranteed traffic rates are sampled over a number of sampling periods of length \( S \). At the end of each measurement window, the rates \( \nu_G^m \) and \( \nu_P^m \) are updated with the maximal rate sampled over the past measurement window (if an individual sample exceeds the current value of \( \nu_G^m \) or \( \nu_P^m \), then the current value is immediately updated, i.e., before the end of the measurement window). The packet delay is updated with the maximal packet delay measured during the past measurement window.

### B. An effective bandwidth based algorithm

The effective bandwidth (EB) based algorithm follows an approach similar to Jamin’s, in that it uses a conservative static AC for the guaranteed traffic and an efficient MBAC for the predictive traffic. However, the admission criteria of the present approach relies on the concept of effective bandwidths rather than delay bounds. An example of an EB-based method is presented in Brichet and Simonian [2], where is focused on the situation of a single (real-time) traffic type.

We adopted the approach of Elwalid, Mitra and Wentworth (EMW) [10] as the basis for the computation of the effective bandwidths of the flows in the MBAC algorithm. The EMW effective bandwidth formula has been developed for static AC and uses the traffic contract (peak rate, burst size and sustained rate) as input. We only use the full traffic contract to determine the effective bandwidth of a newly arriving flow in order to obtain a 'worst case value'. For the computation of the effective bandwidths of existing flows the sustained rate is replaced by an estimate of the actual mean rate obtained from on-line traffic.
In Jamin's algorithm [16] the acceptance criteria consist of two sets: one set checks whether enough resources are available and the other one is concerned with the predefined delay bound for predictive traffic. The EB-based algorithm only consists of criteria of the former kind. The delay bound for predictive traffic is enforced in the computation of the effective bandwidths by exploiting the special form of the EMW formulas in the following way. The buffer size $B$ appears in the EMW formulas only as the numerator of the ratio of $B$ and the available bandwidth (see e.g. Formula (15) in Elwalid et al. [10]), which is exactly the maximum delay a packet will experience. So, if we choose the buffer size for predictive traffic in our effective bandwidth computations equal to the product of the delay bound $D_P$ of predictive traffic and the available link bandwidth, the delay will not exceed $D_P$. The bandwidth available for the predictive traffic flows is estimated to be $C - \nu^P_m$, where $\nu^P_m$ denotes the sum of the measured rates of the guaranteed traffic flows.

The admission rules. Define $e[b, r, p, B, C, \epsilon]$ as the EMW effective bandwidth of a flow with sustained rate $r$, peak rate $p$ and bucket depth $b$, for given link rate $C$, buffer size $B$ and maximum loss ratio $\epsilon$. Denote the measured (mean) rate of an individual guaranteed (predictive) flow $i$ by $\nu^G_m$ ($\nu^P_m$), $B_G$ is the buffer size for guaranteed traffic. The following decision criteria are used in the EB-based algorithm.

* An arriving guaranteed flow characterized by sustained rate $r$, peak rate $p$ and bucket depth $b$, requesting guaranteed service is admitted if the sum of the new flow's requested rate and the reserved bandwidth of the existing guaranteed flows does not exceed the targeted utilization level:

$$uC > r + \nu^G,$$

and if, in addition, the sum of the 'measured' effective bandwidths of the existing flows and the arriving guaranteed flow does not exceed the link bandwidth available (at the same time taking care of the delay bound for predictive traffic):

$$C > e[b, r, p, D_P(C - \nu^G_m), C - \nu^G_m, \epsilon]$$
$$+ \sum_{i \in G} e[i, \nu^G_m, p, B_G, C, \epsilon]$$
$$+ \sum_{i \in P} e[i, \nu^P_m, p, D_P(C - \nu^G_m), C - \nu^G_m, \epsilon].$$

It should be emphasized that this algorithm has a strong 'certainty equivalent character'; for instance the term $C - \nu^G_m - r$ suggests that this amount of bandwidth is available all the time, which is definitely not the case.

The measurement process. To determine an appropriate estimate of the mean rate of an existing flow (which replaces the sustained rate in the effective bandwidth formula), we use a similar measurement process as in Jamin et al. [16], with parameters $S$ and $T$. This is done to make the results of the two MBACs comparable. Note that (in contrast to [16]) the EB-based method does not need packet delay measurements. On the other hand, the EB-based method does require measurements on all the individual flows, whereas the Jamin algorithm only needs to measure the aggregate behavior of both predictive and guaranteed traffic.

IV. NUMERICAL COMPARISON OF THE MBAC ALGORITHMS

We will now present and discuss the results of a simulation study on the performance of the two MBAC algorithms described in the previous subsection. The main goals of the simulations are

* to check the effort needed to find appropriate values for the measurement parameters; secondary aims are to compare the performance of the two MBAC approaches, and to get some first insights into the impact of traffic characteristics on their performance (Section IV.A).
* to assess the efficiency gain achievable by MBAC (Section IV.B).
A. Simulation scenario with artificial traffic

In the simulations we have used homogeneous on/off flows for both the predictive traffic and the guaranteed traffic. During the on-periods data packets are sent at a fixed rate. To capture heavy-tailed phenomena we have not only considered the case of exponentially distributed on- and off-times, but also on-and off time distributions with a heavy tail (in particular the Pareto distribution with finite mean and infinite variance). The flow holding times have an exponential distribution. The flow arrival rates are the same for the guaranteed and the predictive class and are that high, that the system can be considered to be more or less 'saturated' at flow level. Consequently, the system always operates about its maximal utilization level.

In our simulation experiments the traffic flows were not controlled by a token bucket filter, where the algorithms do assume such a filter (see Section III). To adapt to the situation with token bucket, we have chosen in both MBAC algorithms the sustained rate of each flow equal to its peak rate (on-rate) and the maximum burst size equal to 1 packet. Note, that since the peak rate and the actual mean rate are not equal (where peak rate and sustained rate are), bursts can typically be larger than 1 packet. Consequently, using effective bandwidth formulas with measured mean rates instead of sustained rates would lead to under-estimating the effective bandwidth. This will result in admitting too many flows and consequently huge delays or losses. The algorithm of [16] has the same problem (as they replace the sustained rate in the delay bound of [20] by a measured value, and leave the burst size unchanged), although the authors do not mention it explicitly. They neutralize it by using a conservative estimate of the mean. We use the same approach for EB-based algorithm.

The main performance measures of the MBAC algorithms are the achieved link utilization and the fraction of packet delay bound violations, for both the guaranteed traffic and the predictive traffic. In our simulations, the registration of these performance measures starts at the epoch of the first rejected flow. Thus, we gain insight into the performance of our system when it is in a kind of 'stationary state' (due to the 'saturated' character of the system).

Simulation results. We have performed an extensive number of simulation experiments with different values of the measurement parameters $T$ and $S$. We have tuned these parameters such that the link utilization was maximized, while the QoS requirements were still satisfied. Tables 1 and 2 below show some of the results.

Clearly, more simulations should be done (e.g., simulations with flows that have other traffic characteristics than on/off, heterogeneous traffic flows, etc.) to be able to draw more stable conclusions. However, the following observations are certainly worth mentioning:

- No monotonous behaviour could be discovered for $S$. $S$ should not be too large, because otherwise every measurement would give about the mean rate. Consequently then the burstiness is not taken into account appropriately. Regarding $T$ it can be noticed that the total utilization percentage decreases in $T$. This can be explained easily: if $T$ increases, a higher value for the maximum usage rate becomes more likely, which causes the algorithm to act more defensively. In particular, the efficiency of the EB-based algorithm becomes very poor when $T$ is chosen too large. This is due to the fact that, for large $T$ (e.g., $T > 10S$ in our traffic scenario), the measurement process yields for almost each flow a measured 'mean rate' which is equal to the peak rate; in that case the effective bandwidths of these flows will be set too high.

### Table I

<table>
<thead>
<tr>
<th>$T$</th>
<th>average delay</th>
<th>average delay</th>
<th>utilization percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$4S$</td>
<td>1.42</td>
<td>1.43</td>
<td>64.7%</td>
</tr>
<tr>
<td>$6S$</td>
<td>1.40</td>
<td>1.43</td>
<td>69.4%</td>
</tr>
<tr>
<td>$7S$</td>
<td>1.38</td>
<td>1.43</td>
<td>60.5%</td>
</tr>
<tr>
<td>$8S$</td>
<td>1.38</td>
<td>1.43</td>
<td>62.3%</td>
</tr>
<tr>
<td>$9S$</td>
<td>1.34</td>
<td>1.40</td>
<td>59.7%</td>
</tr>
</tbody>
</table>

**Table I**

Results for 'Jamin', for exponential-on/off (left column) and Pareto-on/exponential-off (right columns). Delay given in ms; target delay for predictive traffic is 15 ms, to be exceeded by no more than 1% of the packets.

### Table II

<table>
<thead>
<tr>
<th>$T$</th>
<th>average delay</th>
<th>average delay</th>
<th>utilization percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$4S$</td>
<td>1.37</td>
<td>1.49</td>
<td>60.5%</td>
</tr>
<tr>
<td>$6S$</td>
<td>1.37</td>
<td>1.41</td>
<td>59.8%</td>
</tr>
<tr>
<td>$7S$</td>
<td>1.36</td>
<td>1.41</td>
<td>56.9%</td>
</tr>
<tr>
<td>$8S$</td>
<td>1.33</td>
<td>1.41</td>
<td>56.2%</td>
</tr>
</tbody>
</table>

**Table II**

Results for 'EB-based', for exponential-on/off (left columns) and Pareto-on/exponential-off (right columns). Delay given in ms; target delay for predictive traffic is 10 ms, to be exceeded by no more than 1% of the packets.
equal to the peak rate and, hence, AC will be based on peak rate allocation. Jamin’s algorithm is less sensitive to values of T that are too large, because it is based on rate measurements of aggregated traffic streams.

- Values for S in the neighborhood of the average on- and off-time and T in between 4S and 10S give satisfactory results for both MBAC algorithms (for our specific traffic scenario).
- The optimal choice of measurement parameters (T and S) depends only slightly on the distributions of the lengths of the on- and off-periods. As noticed, the mean lengths are significant!
- The two algorithms show more or less the same performance. Roughly speaking, one could say that Jamin’s algorithm is slightly more efficient than the EB-based algorithm, but the EB-based algorithm is slightly more ‘reliable’.
- Both MBAC algorithms show higher utilization for the case with heavy tailed traffic than for exponential traffic. A possible explanation for this observation might be that in the case of heavy tailed streams the rate measurements are more ‘predictive’ (i.e., provide more information about the behavior of the traffic stream in the near future) than in the case of exponential traffic. Note, that the heavy tail of the on-time distribution (with finite mean and infinite variance) gives rise to the long range dependence (LRD) phenomenon often observed in traces of network traffic [18]. We return to this remarkable fact in Section IV.C.

The last three observations are illustrated by the simulation results listed in the tables above.

### B. Comparison between static AC and MBAC (with traces of real traffic)

One of the major goals of MBAC was to achieve a higher network utilization than would be possible with static AC. Here we compare the two described MBACs with static AC. This is done under practical traffic conditions, i.e., we have used traces of real traffic streams in our simulation programs. These traces were obtained from measurements on Internet traffic transported over an ATM VP and consisted each of a large number of aggregated TCP flows. The figures below show results of the simulations, in particular the utilizations achieved by the EB-based algorithm and Jamin’s MBAC respectively.

The QoS objectives used (and achieved) in the simulations were as follows: a delay bound of 0.05 seconds for predictive traffic, which may be exceeded by about 1% of the packets; practically no queueing delay for the guaranteed packets. It appeared that the impact of the choice of the measurement parameters S and T was not very large, for both MBACs. S equal to about two or three times a packet transmission time and T equal to 3S or 4S yielded the optimal results shown in the figures: a utilisation of about 70%. Note that the performance for both MBACs is comparable.

We have also compared the link utilisation achieved by the MBACs with the utilization that is obtained when static AC would have been applied. For that purpose, we have first determined the ‘best fitting’ SBR traffic contracts for each of the traces and next, based on these traffic contracts, we have determined the corresponding EMW-effective bandwidths (EMW with statistical multiplexing [10]). It turned out that this static AC yields a maximum link utilisation of only 20% to 30% with the traces used in the simulations. This illustrates the tremendous efficiency gain that can be achieved by the application of MBAC.

![Figure 2: Link utilization achieved by MBACs and by static admission control.](image)
CONCLUSIONS AND OPEN PROBLEMS

We end our paper by stating our most important conclusions. We then list the major directions for future research.

A. Conclusions

For an integrated network it is desirable to combine both 'traditional' static AC for guaranteed services (strict QoS requirements) and MBAC for predictive services ('loose' QoS requirements). In conjunction with appropriate scheduling mechanisms (like Weighted Fair Queueing), MBAC is able to exploit the bandwidth reserved for but not used by the guaranteed flows. In this sense both types of ACs are complementary. It should be stressed that the use of MBAC does not exclude strict guarantees for a premium class. This model fits well into the framework of ATM, as well as intserv/IP [23] and diffserv/IP, cf. [1], [22].

The paper addresses a number of features regarding MBAC. As we stressed, MBAC intends to cope with important drawbacks of traditional AC policies:

- a poor efficiency of the traditional traffic descriptor based AC. The reason for this is that a three parameter descriptor can only provide a loose envelope of the real traffic profile (which may even have an instationary nature);
- the user's lack of insight into his traffic profile, such that it is hard to select good traffic parameters.

In this paper we particularly focused on whether the implementation of MBACs is feasible at all, and if yes, what typical efficiency gain can be achieved.

We considered two MBAC schemes from the literature. Both algorithms are of certainty equivalent (CE) type: they are basically static ACs with measured input. The first class is based on effective bandwidths, with measured input parameters; the other on end-to-end delay bounds of regulated traffic. Due to their CE character the methods are inherently non-conservative; they are made safe by using 'pessimistic' measurement procedures. The performance of the MBACs critically relies on the choice of a number of 'measurement parameters' (like measurement intervals, etc.). By simulation, we show that it is tractable to find an appropriate value for them, with a minimal amount of a priori knowledge of the traffic pattern. These settings do not heavily depend on the stochastic properties of the traffic offered. In particu-

C. An analytical model

Triggered by the simulation results we have also performed an analytical study to get more insight into the impact of the on-time distribution on the performance of MBAC, see Mandjes and Van Uitert [19]. To be more precise, we considered the transient of the aggregate rate process of a large number of homogeneous on/off traffic flows, under the assumption that no new calls are admitted after time 0. Given (e.g., from a traffic measurement) that K flows are in the on-state and N-K flows in the off-state at time 0, the probability $P_{N,K}(t)$ that the aggregate rate reaches a certain critical threshold $C$ (e.g. link rate) at time $t$ is computed.

Typically, the probability $P_{N,K}(t)$ is extremely small for small $t$, as it is very unlikely that the aggregate makes a 'jump' in a very short time interval. Then it increases, and at some time $t = t^*$, $P_{N,K}(t)$ reaches a maximum. Then (due to the departure of more and more flows) $P_{N,K}(t)$ starts to decrease to zero. From numerical results we have found, that in the case of exponential on-off flows $P_{N,K}(t^*)$ is considerably larger than in the case of on/off flows with heavy tailed on-and off-times. This supports the above observation that the MBAC performs better for Pareto than exponential on/off traffic. This phenomenon is illustrated by the results depicted in Figure 3.

Another interesting result from the analytical study is that $P_{N,K}(t^*)$ hardly depends on the flow holding time distribution (only its mean is important). For more details, we refer to [19].

V. CONCLUSIONS AND OPEN PROBLEMS

We end our paper by stating our most important conclusions. We then list the major directions for future research.

A. Conclusions

For an integrated network it is desirable to combine both 'traditional' static AC for guaranteed services (strict QoS requirements) and MBAC for predictive services ('loose' QoS requirements). In conjunction with appropriate scheduling mechanisms (like Weighted Fair Queueing), MBAC is able to exploit the bandwidth reserved for but not used by the guaranteed flows. In this sense both types of ACs are complementary. It should be stressed that the use of MBAC does not exclude strict guarantees for a premium class. This model fits well into the framework of ATM, as well as intserv/IP [23] and diffserv/IP, cf. [1], [22].

The paper addresses a number of features regarding MBAC. As we stressed, MBAC intends to cope with important drawbacks of traditional AC policies:

- a poor efficiency of the traditional traffic descriptor based AC. The reason for this is that a three parameter descriptor can only provide a loose envelope of the real traffic profile (which may even have an instationary nature);
- the user's lack of insight into his traffic profile, such that it is hard to select good traffic parameters.

In this paper we particularly focused on whether the implementation of MBACs is feasible at all, and if yes, what typical efficiency gain can be achieved.

We considered two MBAC schemes from the literature. Both algorithms are of certainty equivalent (CE) type: they are basically static ACs with measured input. The first class is based on effective bandwidths, with measured input parameters; the other on end-to-end delay bounds of regulated traffic. Due to their CE character the methods are inherently non-conservative; they are made safe by using 'pessimistic' measurement procedures. The performance of the MBACs critically relies on the choice of a number of 'measurement parameters' (like measurement intervals, etc.). By simulation, we show that it is tractable to find an appropriate value for them, with a minimal amount of a priori knowledge of the traffic pattern. These settings do not heavily depend on the stochastic properties of the traffic offered. In particu-
lar, (i) heavy tailed bursts showed even better utilization than exponential bursts, and (ii) heavy tailed call durations are not expected to do much harm. These results are supported by the theoretical study [19].

The performances of both implemented MBACs (once the parameters are tuned) roughly coincide. This is in line with the conjecture phrased in [17]: every (well-tuned) MBAC has essentially the same packet performance/call performance trade-off (see Section II). However, MBACs differ in the measurement effort required. For instance, an advantage of the approach of Jamin et al. [16] is that it only needs measurements of the aggregate traffic stream, rather than individual flow measurements. A disadvantage of that algorithm is that it requires delay measurements.

In order to assess the efficiency gain in realistic situations, we did experiments with a trace of TCP/IP traffic transported over an ATM virtual path. We identified very significant gains (compared to static AC), but we emphasize that more experiments are needed (particularly with traffic stemming from different applications). A general conclusion from the simulation experiments is that MBAC copes very well with possible instationary sources.

B. Claims and open problems

1. Effect of randomness of the measurements. In this paper we demonstrated that it is feasible to implement a certainty equivalent (CE) method (with a ‘compensating’ measurement mechanism). However, these methods remain quite heuristic, as no guidelines are provided for the tuning of the measurement parameters. A little exaggerated, a CE approach can be viewed as a shift of the problem: instead of fitting a statistical model, we have to find the parameters of the measurement procedure. It was proposed in [3] to dynamically update these tuning parameters; we wonder whether such a procedure is feasible in a real network setting. Due to the lack of transparency of CE methods, we advocate further investigations on the effect of the uncertainty of the measured value, in the spirit of [9], [13], [14].

2. Scalability of measurement procedures. We adhere to the point of view of [17] that (after tuning of the measurement parameters) any MBAC has roughly the same performance (in terms of the trade off between blocking and packet loss). However, still the algorithms are not equivalent, as they are also characterized by the measurement effort required. In the algorithm of [16] even performance measurements (delay) are required; the EB-based algorithm needs per flow measurements and is therefore not scalable. From an implementation point of view, it would be preferable to make only aggregate measurements. A discussion on this issue is given in [14]. The authors of [14] are in favor of aggregate measurements, where the correlations between the measurements are taken into account by using filtering techniques. We propose to extend this to a more systematic assessment of the impact of the assumptions (regarding traffic characteristics, in particular the role of the exponential holding times) and the performance under non-stationary traffic conditions.

3. Theoretical investigations. Further, we are in favor of more theoretical investigations of the sensitivity of transient probabilities. Duffield and Whitt [8] considered the ‘recovery time’: if the MBAC makes a wrong decision, what time does it take to recover? Mandjes and Van Uitert [19] describe (for given traffic characteristics) how well a load measurement predicts future load. The performance of the MBAC is essentially the combination of these two effects; this has not been examined so far.

Also, our simulation studies showed that long tailed traffic does not negatively affect the performance of the MBAC; a more thorough insight into the reasons of this phenomenon would contribute to a better understanding of the impact of traffic characteristics on MBAC performance.

4. Importance of testing with real traces. We stress the importance of the use of real data for testing algorithms. The aim of MBAC is to exploit the gap between contract and real profile, and the only way to verify the efficiency gain is by using traces. Tests based on artificial sources (e.g., on-off sources) can be misleading, as these sources cannot capture the gain achievable by MBAC in case of instationary traffic. In principle, one can always find traffic scenarios that are ‘hard’ for MBACs, but these scenarios may be not representative for real network traffic.

Acknowledgment. Miranda van Uitert wrote the MBAC simulation programs.

REFERENCES


A general approach to admission control


