MBAC: Impact of the Measurement Error on Key Performance Issues

Anne Nevin, Peder J. Emstad, and Yuming Jiang

Centre for Quantifiable Quality of Service in Communication Systems (Q2S)*,
Norwegian University of Science and Technology (NTNU), Trondheim, Norway
{anne.nevin,peder.emstad}@q2s.ntnu.no

Abstract. In Measurement Based Admission Control (MBAC), the decision of accepting or rejecting a new flow is based on measurements of the current traffic situation. Since MBAC relies on measurements, an in-depth understanding of the measurement error and how it is affected by the underlying traffic is vital for the design of a robust MBAC. In this work, we study how the measurement error impacts the admission decision, in terms of *false rejections* and *false acceptances*, and the consequence this has for the MBAC performance. A slack in bandwidth must be added to reduce the probability of false acceptance. When determining the size of this slack, the service provider is confronted with the trade-off between maximizing useful traffic and reducing useless traffic. We show how the system can be provisioned to meet a predefined performance criteria.

1 Introduction

Measurement Based Admission Control (MBAC) has for a long time been recognized as a promising solution for providing statistical Quality of Service(QoS) guarantees in packet switched networks. An MBAC does not require an *a priori* source characterization which in many cases may be difficult or impossible to attain. Instead, MBAC uses measurements to capture the behavior of existing flows and uses this information together with some coarse knowledge of a new flow, when making an admission decision for this requesting flow. The problem with measurements is that they include an error, and the size of this error depends on flow characteristics and the length of the observation window. The errors creates uncertainties that again will effect the admission decisions in terms of *false rejections* and *false acceptances.* False rejections are flows that are rejected when they should have been accepted and false acceptances are flows that are accepted when they should have been rejected. For the service provider, false rejections translates into a decrease in utilization and for the end user, false admissions means that the QoS of the flow can no longer be guaranteed. Basing admission

* "Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence" appointed by The Research Council of Norway, funded by the Research Council, NTNU and UNINETT. http://www.q2s.ntnu.no.

on measurements clearly requires the understanding of the measurement error and how this impacts the performance of MBAC.

Consider a link where the maximum allowable mean rate is uc and the main task is to keep the average workload of the link at or below this level. An MBAC is put in place which uses measurements to find an estimate $\hat{R}$ of the mean aggregate rate of accepted flows. When a new flow arrives, with mean rate ξ, it will be accepted if:

$$\hat{R} + \xi \leq uc. \tag{1}$$

With MBAC, care must be taken, since $\hat{R}$ is not the true mean rate and includes an error, which will cause *false admissions* and *false rejections*. The measurements are improved when they are taken over a longer measurement window. However, flows leaving within the window results in flawed estimates, thus the flow lifetimes set an upper limit for the window size. Given this window size, how confident can be be this is not a false acceptance? Is this good enough? If the answer is no, the reserved bandwidth for the flows uc, must be reduced by some slack to make up for the measurement uncertainty. But how large should this slack in bandwidth be?

There is a tradeoff between rejecting too many flows thus wasting resources, and accepting too many flows resulting in QoS violations. In this work we study how the measurement errors and flow dynamics impact the performance of MBAC in terms of proper performances measures. A simple example shows how the system can be provisioned with a predefined performance criteria.

The focus in the literature has been on finding MBAC algorithms that maximize utilization while providing QoS which basically implies the determination of uc (see [1] and [2] for an overview). After studying various MBAC algorithms, [1] concluded that all algorithms have nearly the same performance with similar deviations from the *ideal* behavior. We claim that this conclusions is not surprising and is primarily due to measurement errors. A deeper analytical understanding of the measurement process and its error has been sought in [3] and [4]. Our work differs significantly from previous work in that correlation characteristics within flows are included and we find how the uncertainty in the measurements vary with the length of the observation window. This work is based on previous work [5] and is part of a methodology and design of an analytical framework for analyzing measurement error.

The reminder of the paper is organized as follows. First the system model is introduced in Section 2. Section 3 details the rate level measurements and Section 4 introduces the flow level framework and defines the performance measures. Provisioning is discussed in Section 5 and follows up with a case study in Section 6, before the conclusion is given in Section 7.

2 System Model

Flows compete for a limited resource, a network link of capacity c, controlled by MBAC. The flow's QoS requirement can be guaranteed as long as the average aggregate rate is at or below uc, where u, $0 < u < 1$ is a tuning parameter. An optimal value for u depends on flow characteristics. In this work, u is assumed a given constant and a discussion around its optimal settings is out of scope. The MBAC works as follows: The MBAC estimates the average aggregate rate $\hat{R}$ based on measurements taken periodically every measurement window of size w. When a new flow arrives, it will be accepted or rejected at the start of the next measurement period according to (1). Additional flows arriving within the same window will be rejected and lost.

Each flow is a stationary rate process with mean ξ and covariance $\rho(\tau)$. A mixing of flow classes will cause increased complexity for the MBAC algorithm and also the measurement process. To simplify, only the homogenous case where flows belong to the same class will be considered. With this assumption, the *system state* N can be specified by the current number of flows. The maximum number of flows the system can handle, is thus $n_{max} = uc/\xi$ and the MBAC algorithm can be written:

$$\hat{R} + \xi \leq n_{max}\xi \tag{2}$$

It is natural to separate the timescale into the rate level where the measurements are done and the flow level where the admission decision is made. The measurements are taken by means of continuous observation and will be discussed next.

3 Rate Level: Measurements and Measurement Error

The flow rate process $X_i(t)$ is observed continuously over the window and an estimate of the mean, $\hat{X}_i$ is then: [5]:

$$\hat{X}_i = \frac{1}{w}\int_0^w X_i(t)dt \tag{3}$$

This measured statistic varies with the window size according to [5]:

$$\zeta^2(w) = \frac{2}{w^2}\int_0^w (w-t)\rho(t)dt \tag{4}$$

The aggregate rate process is also a stationary rate process. Conditioned on being in state n, the aggregate mean is $n\xi$. An estimate of the aggregate mean is given by:

$$\hat{R} = \sum_{i=1}^{n}\frac{1}{w}\int_0^w X_i(t)dt \tag{5}$$

When the number of flows is large (say $n > 30$), the sum of the average over the flows will be close to a normal distribution thus $\hat{R} \sim \mathcal{N}(n\xi, n\zeta^2(w))$. This assumption will be made here.

The accuracy of this measurement can then be described by the $1-\varepsilon$ confidence interval: $\hat{R} - z_{\frac{\varepsilon}{2}}\sqrt{n}\zeta(w) \leq n\xi < \hat{R} + z_{\frac{\varepsilon}{2}}\sqrt{n}\zeta(w)$, where $z_{\frac{\varepsilon}{2}}$ is the $(1-\alpha/2)$ quantile of the normal distribution.

It is intuitive to think that in order to achieve a certain measurement accuracy all that is needed is to increase the window size. However in order for the above estimate to hold, the requirement is that no flows leave during the window, i.e. the aggregate rate process is stationary with a known distribution. Otherwise the actual estimate becomes incorrect. The flow lifetime therefore sets an upper limit for the window size.

The inaccuracy of the measurement translates into uncertainty in the admission decision process done at the flow level and will be described next.

4 Flow Level and Performance Measures

Let new flows arrive following a Poisson process with parameter λ. If the flow is accepted it stays in the system for a lifetime that is negative exponentially distributed with mean $1/\mu$. A flow that is not accepted by the MBAC is lost. The *offered flow load* is the Erlang load [6] denoted by A. This is the average number of simultaneous flows if there is no blocking given by:

$$A = \lambda \cdot E(T_L). \tag{6}$$

With the assumption that $\hat{R} \sim \mathcal{N}(n\xi, n\zeta^2(w))$, if there are i flows in the system, a new arriving flow will be accepted with a probability $q_i = P(\hat{R} + \xi \leq n_{max} \mid N = i)$. We will assume that the arrival rate is such that the probability of more than one flow arrival per window is very small. The lost traffic due to multiple arrivals within the window is thus very small and can be neglected.

The number of flows currently accepted by the MBAC follows a continuous time Markov chain, see Fig. 1 and the probability that there are i flows in the system is:

$$P(i) = \frac{\frac{A^i}{i!}\prod_{x=0}^{i-1} q_x}{\sum_{j=0}^{\infty} \frac{A^j}{j!}\prod_{x=0}^{j-1} q_x} \tag{7}$$

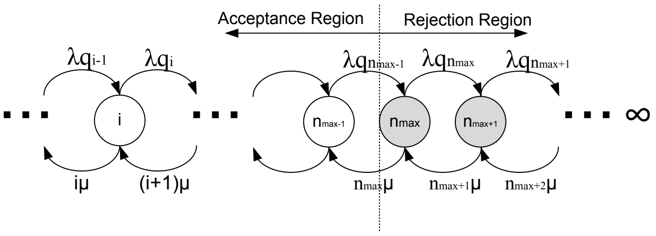


Fig. 1. State diagram of the number of sources accepted by the MBAC

In [7] a similar framework is defined, where the q_is are based on the probability that the instantaneous rate measurements are above a threshold. However, they do not specify how these measurements should be taken. Since the instantaneous rate is a random quantity without a true value, it can not be measured [8]. This is where our work deviates significantly from [7]. We study the measurement error in isolation and the q_is depend on how accurate the measurements are.

Implied by (7) and as also discussed in [7], the distribution $P(i)$ is indeed insensitive to the distribution of flow lifetime and only depends on the expected flow lifetime.

The system state space can be divided into two regions; the *acceptance region*, $N < n_{max}$ and the rejection region, $N \geq n_{max}$ see Fig. 1.

Rejecting a flow when the system is in the acceptance region is a *False rejection* and accepting a flow when the system is in the rejection region is a *False acceptance.*

For assessing the performance and provisioning purposes, we define the following flow level performance measures:

- **Probability of False acceptance,** P_{FAcc}, is the probability that an arriving flow is accepted in the rejection region

$$P_{FAcc} = \sum_{i=n_{max}}^{\infty} P(i) \cdot q_i \tag{8}$$

- **Probability of False rejection,** P_{FRej}, is the probability that an arriving flow is rejected when it should have been accepted

$$P_{FRej} = \sum_{i=1}^{n_{max}-1} (1 - q_i)P(i) \tag{9}$$

- **Blocking probability,** P_B, is the probability that an arriving flow is rejected.

$$P_B = P_{FRej} + P(N \geq n_{max} \cap rejection) = \sum_{i=1}^{\infty}(1 - q_i)P(i) \tag{10}$$

- **Carried useful traffic,** A_{useful} , is the expected number of flows in the acceptance region.

$$A_{useful} = \sum_{i=0}^{n_{max}} iP(i) \tag{11}$$

- **Carried useless traffic,** $A_{useless}$, is the expected number of flows in the rejection region.

$$A_{useless} = \sum_{i=n_{max}+1}^{\infty} iP(i) \tag{12}$$

- **Lost Traffic,** A_{lost} , is the traffic that is blocked from the network

$$A_{lost} = AP_B \tag{13}$$

If there are no measurement errors, the admission controller becomes *ideal*, $\hat{R} = \bar{R}$ and the distribution of flows is then as for the Erlang Loss system. This system only carries useful traffic and arriving flows will experience a *the blocking probability* given by the Erlang B formula.

4.1 Flow load and window size limitations

Our defined framework only considers the flow load $A = \lambda/\mu$. When increasing A, it is indifferent if this is done by increasing the flow lifetime $1/\mu$ or increasing the arrival rate λ. In reality this is not true. Many arrivals within a measurement window will increase the blocking probability since the MBAC only admits at most one flow after a measurement update. Our model neglects this lost traffic and simply assumes that the probability of more than one arrival within a window is very small.

As explained in Section 3, the accuracy of the measurement itself will decrease if flows leave during the measurement window. Clearly, for a constant A, a longer measurement window can be used for long lifetimes (infrequent arrivals) as compared to short lifetimes (frequent arrivals). In addition, what also must be kept in mind is that a large window size will result in longer setup times as an arriving flow must wait for a measurement update before it can be accepted.

When determining a proper window size there are thus three factors to consider: 1) The flow arrival rate, 2) The flow lifetime and, 3) The required connection setup time. For the purpose of this study, the window size will not be considered a design parameter and is the largest window possible while still conforming to the model assumptions.

5 Provisioning

The QoS provided to the flows can only be guaranteed as long as the number of flows is at or below n_{max}, thus admitting more than n_{max} flows should be avoided. If the probability of false acceptance is too high, a safeguard in terms of a slack in bandwidth can be added to make up for the measurement errors.

As in Paper [5], let the safeguard have increments of size size $l\xi, 0 < l < n_{max}$ and the refined admission control algorithm becomes:

$$\hat{R} + \xi \leq \xi n_{max} - l\xi \tag{14}$$

The critical situation arise as soon as the system reaches state n_{max}, where accepting a flow will result in the first false admission. With the condition that the system is in state n_{max} we define the *conditional performance requirement*:

$$P(FAcc \mid N = n_{max}) = P(\hat{R} + \xi \leq \xi n_{max} \mid N = n_{max}) \leq \epsilon \tag{15}$$

where ϵ is termed the *conditional performance target.*

For a given quantile and predefined window size, $P(FAcc \mid N = n_{max})$ can be kept below the target if the number of levels is [5]:

$$l + 1 = \left\lceil \frac{\sqrt{n_{max}}\zeta(w) z_{\varepsilon}}{\xi} \right\rceil . \tag{16}$$

Where z_{ε} is the ε- quantile of the standard normal distribution The resulting l can be used for provisioning the system.

Since the MBAC solely estimates the number of flows through measurements, l is independent of the system state and when there are i flows in the system, a new arriving flow will be accepted by the MBAC, with a probability $q_i = P(\hat{R} + \xi \leq n_{max} - l \mid N = i)$.

The size of l controls the probability of entering the rejection region by forcing the probability distribution towards left (Fig. 1). Obviously, if the slack is too large, the MBAC becomes too pessimistic and resources are wasted unnecessarily. The actual performance can be evaluated by means of the performance measures defined in Section 4. To the customer, the performance measures of interest are the blocking probability and the probability of false acceptance. The service provider seeks to balance the carried useless traffic and carried useful traffic.

For a given flow load the network can be provisioned to meet a desired performance target. But what flow load should be used?

The network should be dimensioned to ensure a small blocking probability under normal loads, say $P_B < 0.01$. At such low loads the probability of entering the rejection region is very small and excellent QoS can be provided to all flows. The problem arise in times of excessive demand. With an *ideal* controller, when the load increases above what is predicted the blocking probability increases to unacceptable high values. However, the QoS to the already admitted flows will not be harmed. With MBAC on the other hand, also the probability of false acceptance and useless traffic increase with increasing loads. Reviewing work in the MBAC literature, it is common practise to test the performance under heavy flow load, resulting in 50% blocking probability(e.g [1]) or infinite load (e.g.[3]).

We do not attempt to answer, exactly what load to use for provisioning purposes. The load must be relatively high, since the main task of MBAC is to preserve QoS to its users when the load exceeds normal values [9]. Obviously at such loads, the normal blocking probability (e.g $P_B = 0.01$) can not be met.

6 Case study using MMRP source models

In this section provisioning to fulfill some predefined performance criteria will be demonstrated with an example.

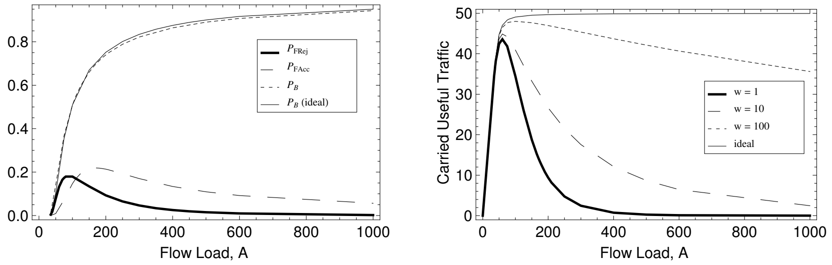
Let the flows be modeled by a two-state Markov modulated rate process (MMRP) which is a simple, yet realistic source model used to model both speech sources and video sources [10]. The MMRP process $X(t) = rI(t)$ where $I(t)$ alternates between states $I = 0$ and $I = 1$ and r is the peak rate. The duration of the 0 and 1 states follows a negative exponential distribution with mean $1/\alpha$ and $1/\beta$ respectively.

The variance of the time average of such a source is [5]:

$$\zeta^2(w) = \frac{2r^2\alpha\beta}{w^2(\alpha+\beta)^3}\left(w - \frac{1}{\alpha+\beta}(1 - e^{-w(\alpha+\beta)})\right) \tag{17}$$

In the following, $\alpha = \beta = r = 2$, and ξ is then 1. The flows have a QoS requirement that can only be guaranteed as long as $N \leq n_{max} = 50$.

First we shall show how the performance measures are impacted by variation in the offered flow load, A and without a slack in the bandwidth (e.i. $l = 0$) Keeping the window size constant at $w = 1$, Fig. 2(a) shows how the performance measures P_{FAcc}, P_{FRej} and P_B are impacted by increasing loads. Low loads will result in negligible false acceptance. Instead false rejections cause a slight increase in blocking probability as compared to the *ideal*. At a load of about $A = 60$, $P_{FAcc} = P_{Frej}$ and then as the load increases P_{FAcc} increases resulting in a slightly lower blocking probability as compared to the *ideal*. As the load increases towards infinity P_{FAcc} becomes zero. The reason for this is that the system moves into the rejection region and will eventually only carry useless traffic. This is illustrated in Fig. 2(b) which shows that as the load increases the carried useful traffic approaches zero. Also shown is that for larger window sizes, the MBAC approaches the *ideal* and the carried useful traffic falls off slower.



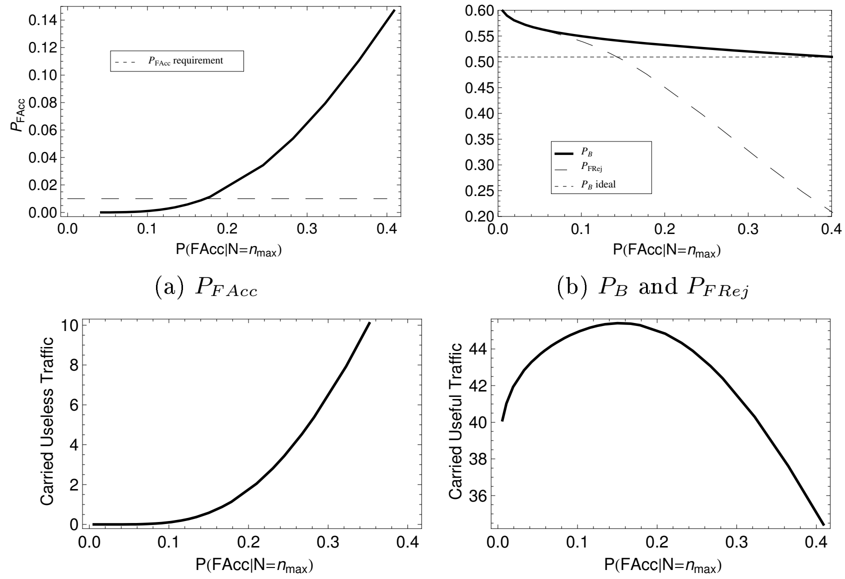
(a) P_{FRej} and P_{FAcc} vs load (b) Carried useful traffic

Fig. 2. Probability of false acceptance as the load increase and the corresponding Carried Useful Traffic

Let the system be provisioned to handle a load of $A = 100$ and let the window size still be $w = 1$. We are interested in finding the required safeguard (number of levels) needed to maximize the carried useful traffic while keeping $P_{FAcc} < 0.01$.

Increasing $P(FAcc \mid N = n_{max})$ translates into a decrease in number of levels where $l = 0$ corresponds to the maximum value $P(FAcc \mid N = n_{max}) = 0.4$. The performance plots shown in Fig. 3(a), 3(b), 3(c), and 3(d) illustrate the tradeoff between blocking and accepting flows when $P(FAcc \mid N = n_{max})$ is varied. Consider first the performance in the light of the customer. To fulfill the requirement of $P_{FAcc} < 0.01$, Fig.3(a), shows that $P(FAcc \mid N = n_{max}) < 0.17$. At this value the blocking probability is about 5% larger than for the *ideal*.

For the service provider, the concern is false rejections and useless traffic. As $P(FAcc \mid N = n_{max})$ increases, false rejections fall off (see Fig. 3(b)) and the useless traffic increases (see Fig.3(c)). Observe in Fig.3(d) that a value $P(FAcc \mid N = n_{max}) = 0.15$ maximizes the carried useful traffic. Reducing the value and the admission controller becomes too strict due to the increase in P_{FRej}. Increasing the value passed this point on the other hand and the admission controller accepts too much useless traffic. In this case, using $P(FAcc \mid N = n_{max}) = 0.15$, also ensures that $P_{FAcc} < 0.01$. The required levels can then be found using (16) and will thus be $l = 4$. Using a higher flow load A and the value $P(FAcc \mid N = n_{max}) = \epsilon$ which maximize the carried traffic will decrease resulting in a larger required safeguard. For example, using the extreme load of $A = 1000$ results in $l = 9$. Decreasing the load, will have the opposite effect, and eventually as the load is reduced further, the maximum will be reached for an ϵ where no safeguard is needed.



(a) P_{FAcc}

(b) P_B and P_{FRej}

(c) Carried useless traffic

(d) Carried useful traffic

Fig. 3. The performance measures as $P(FAcc \mid N = n_{max})$ varies for $A = 100$ and $w = 10$ a)Probability of false acceptance, b) Probability of false rejection and overall blocking probability, c) Carried useless traffic and d) Carried useful traffic

7 Conclusion

This current work gives an in-depth understanding of how measurement uncertainties and flow dynamics impact the MBAC admission decision. An MBAC algorithm, no matter how advanced, is of little use if the measurement errors are not taken into account. These errors translate into uncertainty in the decision process. The degree of uncertainty abates with the length of the observation window. Despite the heavy reliance on measurements, there is in the literature of MBAC, surprisingly very limited work focusing on the impact of the measurement error and how it affects the admission decision.

The probability of false admissions can be reduced by adding a slack in bandwidth. However, if the slack is too large, flows are blocked unnecessarily. With some appropriate performance measures, we showed how the system can be provisioned to meet a predefined performance criteria.

In this work, we assumed that the flows were homogenous, however in a separate work we have developed the analytical tools needed to extend this work to the non-homogenous case.

References

1. L. Breslau, S. Jamin, and S. Shenker, "Comments on the performance of measurement-based admission control algorithms," in *IEEE INFOCOM*, 2000.
2. A.W.Moore, "Measurement-based management of network resources," Technical Report, University of Cambridg, Cambridge CB3 OFD, United Kingdom, April 2002.
3. M. Grossglauser and D. N. Tse, "A time-scale decomposition approach to measurement-based admission control," *IEEE/ACM Trans. Networking*, vol. 11, no. 4, pp. 550–563, Aug. 2003.
4. Z. Dziong, M. Juda, and L. G. Mason, "A framework for bandwidth management in ATM networks - aggregate equivalent bandwidth estimation approach," *IEEE/ACM Trans. Networking*, vol. 5, no. 1, pp. 134–147, Feb. 1997.
5. A. Nevin, P. Emstad, Y. Jiang, and G. Hu, "Quantifying the uncertainty in measurements for mbac," in *Proc. EUNICE 2009*, 2009.
6. V. B. Iversen, "Teletraffic engineering and network planning," May 2006, course Textbook.
7. P. K. R.J. Gibbens, F.P. Kelly, "A decision-theoretic approach to call admission control in atm networks," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 13, no. 6, pp. 1101–1114, Aug. 1995.
8. S. G. Rabinovich, *Measurement Errors and Uncertaintees*, 2nd ed. Springer-Verlag New York, 2005.
9. J.W.Roberts, "Internet traffic, qos and prising," in *Proceedings of the IEEE*, vol. 92, no. 9, 2004.
10. M.Schwartz, *Broadband integrated networks*, P.Becker, Ed. Prentice Hall, 1996.