

MBAC: The Measurement Error with Non-Homogeneous Flows

Anne Nevin, Peder J. Emstad

Centre for Quantifiable Quality of Service in Communication Systems (Q2S) ¹

Norwegian University of Science and Technology, Trondheim, Norway

Email: {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. 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. The consequence of the measurement error is that flows are accepted in error, a *false acceptance* and rejected in error, a *false rejection*. In this paper, the focus is on the prevention of false acceptance as the consequence is QoS violations and a service that may be of little use to the customers. Non-homogeneous flows cause increased complexity for the MBAC algorithm and also for the measurement process. The concept of *similar flows* is introduced, which is a restriction to simplify the analytical expressions in a non-homogeneous flow environment. This work differs significantly from previous work in the literature in that the measurement error is characterized as it abates with the length of the measurement window.

I. 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 that 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 the requesting flow.

A new flow should only be accepted if the admission controller can say yes to the following basic admission criteria:

- Are there sufficient resources to meet the QoS requirement of the arriving flow?
- If the flow is accepted will the QoS of the already accepted flows still be met?

The problem with MBAC is that measurements are unavoidably inaccurate. This imperfection creates uncertainties which affect the MBAC decision process. Flows will be accepted when they should have been rejected, *false acceptance*, and rejected when they should have been accepted, *false rejections*. Clearly, by answering yes to the above questions when the answer should have been no will put all the flows at risk of QoS violations. When the QoS requirement is not fulfilled, the service may be of little use to the customers, thus an in-depth

understanding of the measurements themselves and how they are affected by the underlying traffic is vital for the design of a robust MBAC.

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 we be that this is not a false acceptance? To make up for the measurement error, the reserved bandwidth for the flows must be reduced by some slack. But how large should this slack in bandwidth be?

The objective of this paper is to quantify the probability of false acceptance due to the uncertainty of the measured average rate. Paper [1] analyzes the measurement error in a system with homogeneous flows. Non-homogeneous flows cause increased complexity for the MBAC algorithm and also the measurement process. However, the assumption of flows being homogeneous is very restrictive, since there are increasing types of traffic/applications in the Internet. Even flows belonging to the same application e.g video streaming or VoIP, may not be strictly homogeneous. Hence, that the same traffic type constitute of homogeneous flows is not a realistic assumption. In this work, a (much) relaxed assumption is adopted, i.e. flows can be grouped as being a class of "similar flows" if they share a common correlation structure. The concept of *similar flows*, is a restriction to simplify the analytical expressions in a non-homogeneous flow environment.

For the performance analysis the focus will be on the probability of false acceptance without considering flow arrivals and departures. A separate work [2] includes flow dynamics and how this impacts some performance measures of MBAC.

The focus in the literature has been on finding MBAC algorithms that maximize utilization while providing QoS (see [3] and [4] for an overview). However, very little work has been on understanding the statistical nature of the measured parameters themselves [5]. A proper setting of the length of the measurement period has been of general concern in the MBAC literature. Based on simulations, the only conclusion that can be drawn is that different settings is needed for different traffic scenarios. A deeper analytical understanding of the measurement process and its error has been sought in [5] and [6]. Our work differs significantly from previous work in that we include correlation characteristics within flows and we find how the uncertainty in the measurements vary with

¹"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>

the length of the observation window. This work is based on previous work [1] and is part of a methodology and design of an analytical framework for analyzing measurement error.

This chapter is organized as follows: Before the analysis of the measurement error for non-homogeneous flows are given, analytical means to describe the flow rate process and flow mix is needed. Similar flows will be used to describe the rate process of non-homogeneous flows and this concept is introduced in Section II. To describe the flow mix, the multi-dimensional knapsack model is given in Section IV, and the measurement error is characterized in Section V. Section VI follows up with a case study to demonstrate the use of the similar flows concept. A conclusion is given in Section VII.

II. SYSTEM MODEL AND THE CONCEPT OF SIMILAR FLOWS

Flows arrive to a single node network link of capacity c . The network is flow-aware as described in [7] and [8], where user-defined flows are identified on the fly. In a flow-aware network, admission control is local to a particular network link, where local traffic and service information can be easily obtained.

The flows are non-homogeneous and the concept of traffic classes, often used in the literature ([9], [10]) will be used to distinguish between different types of flows. Specifically, let there be k classes of flows, where the members of a given class are those with the same value of the traffic parameters. In our definition flows belong to the same particular class i , if they have the same *distribution of flow life time* and *rate dynamics*.

All flows are taken to be independent and at the rate level, it is assumed that the flow rate process $K_i(t)$ is a stationary rate process described by its mean ξ_i and auto covariance function.

The flows have a QoS requirement which can only 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.

An MBAC is put in place to control access to this link and prevent the average aggregate rate from exceeding its upper limit, uc . Upon a flow arrival, it is assumed that the MBAC can distinguish between classes of flows but has no knowledge regarding flow departures.

To assess information about the state of the system, the MBAC measures the average aggregate rate \hat{R} based on observations of the aggregate rate $R(t)$ over a measurement window of size w . This measurement replaces the measurement taken in the previous window. A new arriving flow carries with it a bandwidth requirement ξ_i which is fed to the MBAC. When a new flow arrives, it will be accepted if:

$$\hat{R} + \xi_i \leq uc. \quad (1)$$

Otherwise, the flow will be rejected. Additional flows arriving within the measurement window are also denied admission. Most algorithms in the MBAC literature uses the

peak rate of the incoming flow instead of the mean rate ξ_i . Assuming peak rate of the arriving flow is perhaps more realistic but makes the MBAC more pessimistic. We use the mean rate so that the MBAC can be directly mapped to an ideal admission controller which does not base its decision on measurements. As will be seen later, the analytical analysis will also fit the peak rate assumption.

The measured value \hat{R} deviates from the true value \bar{R} due to the *measurement error*, $\delta = \hat{R} - \bar{R}$. In reality, the true value of the measurement is always unknown and in order to describe the measurement error and the consequence it will have on MBAC performance, this must be investigated theoretically.

When performing the analytical analysis, the underlying parameters are disclosed and the measurement error can be stated up front. Let the vector $\mathbf{N} = \mathbf{n} = (n_1, n_2, \dots, n_k)$ describe the current state of the system, the number of accepted flows from each class. Conditioned on the system being in a particular state \mathbf{n} and the mean of the aggregate process is:

$$\xi_{\mathbf{n}} = \sum_{i=1}^k n_i \xi_i \quad (2)$$

Multiple classes of flows complicate the analytical error analysis. However, assuming that flows are homogeneous (i.e. belongs to the same class) is very restrictive, even if flows are of same type e.g only video applications. The concept of *similar flows* which is a special case of non-homogeneous flows, is introduced to simplify this analysis. Flows are said to be *similar* if they obey some restriction on their rates and maximum variance and all have the same auto-correlation function $\Psi(t)$. It is reasonable to assume that a common correlation structure can be found and representative if the number of similar flows is large.

The rate process $K_i(t)$ of a similar flow belonging to class i with mean ξ_i and variance σ_i^2 has the following requirements:

$$\begin{aligned} \kappa &\leq \xi_i \leq r_{max} \\ \sigma_i^2 &\leq \sigma_{max}^2 \\ \frac{cov(K_i(t), K_i(t + \tau))}{\sigma_i^2} &= \Psi(\tau) \end{aligned}$$

where $\kappa > 0$ is a lower bound on the mean rate and σ_{max}^2 is an upper bound on the variance. The maximum number of flows belonging to class i , that can be aggregated on the link is controlled by the mean value ξ_i and the lower bound restriction κ is necessary in that it limits the number of flows and thereby the variance of the aggregated flows.

III. MEASUREMENT PROCESS

Since MBAC has no knowledge about the flows in the system, it relies on an estimate of the mean rate. This section details the method of obtaining the measured statistics and the derivation of the analytical formulas needed for analyzing the measurement error.

The individual flow rate process $K(t)$ is observed every time slot Δ , where X_i is the observation at the end of time

slot i . A measurement window w , consists of m observations of the process, $w = m\Delta$, see Fig. 1.

The rate process has mean ξ and is assumed to be covariance stationary with covariance function $\rho(h)$:

$$\begin{aligned}\rho(h) &= \text{cov}(X_i, X_{i+h}) \\ &= E\{(X_i - E[X_i])(X_{i+h} - E[X_{i+h}])\} \\ &= E(X_i X_{i+h}) - \xi^2\end{aligned}\quad (3)$$

A. Measurement Method 1: Equidistant Sampling

With equidistant sampling, an instant observation of the rate $K(t)$ is taken at every $t = \Delta i$. X_i is the measured rate at the end of time slot i given by $X_i = K(t_i)$. The measured sample $X = X_1, X_2, \dots, X_m$ will be identically distributed but correlated observations, where the X_i s have a sample mean, \bar{X} given by

$$\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i \quad (4)$$

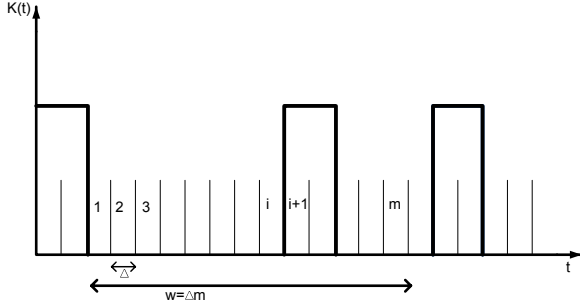


Fig. 1. Rate $K(t)$ vs time

A general expression for the variance of \bar{X} , $\text{Var}(\bar{X})$, is given by [11]:

$$\begin{aligned}\text{Var}(\bar{X}) &= E[(\bar{X} - \xi)^2] \\ &= \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m E[(X_i - \xi)(X_j - \xi)]\end{aligned}\quad (5)$$

and with a covariance stationary process:

$$\text{Var}(\bar{X}) = \frac{1}{m^2} \sum_{h=1-m}^{m-1} (m - |h|) \rho(h) \quad (6)$$

B. Measurement Method 2: Continuous Observation

The best estimate of the mean rate over the window is found by continuous observation. Analytically this is done by letting the sampling rate go towards infinity.

Let now $\Delta \rightarrow 0$ and $m \rightarrow \infty$ keeping the product $m\Delta$ constant such that $t_i = i\Delta \Rightarrow t$ then:

$$\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i \Rightarrow \lim_{\Delta \rightarrow 0 | w = m\Delta} \frac{1}{m} \sum_{i=1}^m X_i = \frac{1}{w} \int_0^w K(t) dt \quad (7)$$

Using limit considerations known from the literature, the variance of the time average, $\zeta^2(w)$ can be found:

$$\begin{aligned}\zeta^2(w) &= \lim_{\Delta \rightarrow 0 | w = m\Delta} \text{Var}(\bar{X}) \\ &= \lim_{\Delta \rightarrow 0 | w = m\Delta} \left(\frac{\Delta}{w}\right)^2 \sum_{i=1}^{m-1} (m - |i|) \rho(t_i) \\ &= \frac{1}{w^2} \int_{-w}^w (w - |t|) \rho(t) dt \\ &= \frac{2}{w^2} \int_0^w (w - t) \rho(t) dt\end{aligned}\quad (8)$$

Note that $\zeta^2(w)$ only depends on the window size and the auto-covariance function $\rho(t)$.

In the remainder the mean rate is always estimated by means of continuous observation.

Conditioned on the system being in a particular state $\mathbf{n} = \{n_1, n_2, \dots, n_k\}$, an estimate of the aggregate mean is given by:

$$\hat{R} = \sum_{i=1}^k \sum_{j=1}^{n_i} \frac{1}{w} \int_0^w K_j(t) \quad (9)$$

The covariance of the aggregate rate is $\sum_{i=1}^k n_i \text{cov}(K_i(t), K_i(t + \tau))$ and inserting into (8), the expression for the variance of the time average of the aggregate is:

$$\zeta_{\mathbf{n}}^2(w) = \frac{2}{w^2} \int_0^w (w - t) \sum_{i=1}^k n_i \text{cov}(K_i(t), K_i(t + \tau)) dt \quad (10)$$

When the number of flows is in the order of a few tens [12] (e.g 30 flows), the sum of the average over the flows will be close to a normal distribution, thus $\hat{R} \sim \mathcal{N}(\xi_{\mathbf{n}}, \zeta_{\mathbf{n}}^2(w))$. This assumption will be made here.

The accuracy of this measurement can then be described by the $1 - \varepsilon$ confidence interval:

$$\hat{R} - z_{\frac{\varepsilon}{2}} \zeta_{\mathbf{n}}(w) \leq \xi_{\mathbf{n}} < \hat{R} + z_{\frac{\varepsilon}{2}} \zeta_{\mathbf{n}}(w)$$

where $z_{\frac{\varepsilon}{2}}$ is the $(1 - \varepsilon/2)$ 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.

IV. IDEAL ADMISSION CONTROLLER AND THE STOCHASTIC KNAPSACK

Before looking into the measurement error, consider a system where the admission controller has perfect knowledge of the aggregate mean rate. In this system, there is no measurements error and \hat{R} is replaced with the true value \bar{R} in (1). This admission controller is referred to as the *ideal* controller. This *ideal* controller will always accept a flow from class i , when

the system is in the *acceptance region*, $\bar{R} \leq uc - \xi_i$. When $\bar{R} > uc - \xi_i$ the system is in the *rejection region* and a flow is always rejected thus for this system \bar{R} will never exceed uc . Fig. 2 gives an illustration of the two class dependent regions.

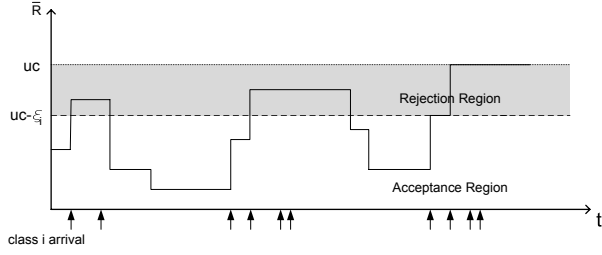


Fig. 2. Illustration of the rejection region and acceptance region for class i

Let new flows belonging to class i arrive following a Poisson process with parameter λ_i . If the flow is accepted it stays in the system for a negative exponentially distributed lifetime with mean $1/\mu_i$. A flow that is not accepted is lost. The *offered flow load* from class i , is the Erlang load [13] denoted by A_i . This is the average number of simultaneous flows if there is no blocking given by:

$$A_i = \frac{\lambda_i}{\mu_i} \quad (11)$$

The system can now be modeled by means of a Stochastic Knapsack [14] and supporting literature for this section can be found in [14], chapter 2.

Convert uc into l_{max} discrete resource units of size ξ , where ξ is the largest common denominator of all ξ_i such that $uc = l_{max}\xi$. Also convert the mean rate ξ_i of class i , into $b_i, 1 < b_i < l_{max}$ units of size ξ . The stochastic knapsack then consists of l_{max} resource units, where a flow from class i will require b_i resources. If there are enough resources available, the flow is accepted and will occupy b_i resource units throughout the duration of the flow.

Conditioned on the system being in a particular state $\mathbf{n} = (n_1, n_2, \dots, n_k)$, the total amount of resources currently in use is given by \mathbf{bn} , where $\mathbf{b} = (b_1, \dots, b_k)$.

$$\mathbf{bn} = \sum_{i=1}^k n_i b_i \quad (12)$$

Define the system state space:

$$\mathcal{S} = \{(n_1, \dots, n_i, \dots, n_k) : \mathbf{bn} \leq l_{max}\} \quad (13)$$

For a class i flow, the acceptance region is the set of states \mathbf{n} where the knapsack will admit a class i flow:

$$A_i = \{\mathbf{n} \in \mathcal{S} : \mathbf{bn} \leq l_{max} - b_i\} \quad (14)$$

For a class i flow, the rejection region is the subset of states where a flow class i will be rejected :

$$\mathcal{Q}_i = \{\mathbf{n} \in \mathcal{S} : l_{max} - b_i < \mathbf{bn}\}$$

Denote the equilibrium state probability $\pi(\mathbf{n})$ as the probability of the system being in state \mathbf{n} . The state probabilities are known to have a product form solution [14]:

$$P(\mathbf{N} = \mathbf{n}) = \pi(\mathbf{n}) = G^{-1} \prod_{i=1}^k \frac{A_i^{n_i}}{n_i!} \quad (15)$$

where G is the normalization constant:

$$G = \sum_{\mathbf{n} \in \mathcal{S}} \prod_{i=1}^k \frac{A_i^{n_i}}{n_i!} \quad (16)$$

Note, that the product form solution is insensitive to the distribution of the flow lifetime and only depends on the mean [14].

Let \mathbf{q}_i be a particular state within the rejection region of class i , $\mathbf{q}_i \in \mathcal{Q}_i$.

Define now the conditional blocking probability $P_{\mathcal{Q}_i}(\mathbf{q}_i)$ as the probability of being in a rejection state \mathbf{q}_i given that the system is in the rejection region for class i .

$$P_{\mathcal{Q}_i}(\mathbf{q}_i) = P(\mathbf{N} = \mathbf{q}_i | \mathbf{N} \in \mathcal{Q}_i) = \frac{\pi(\mathbf{q}_i)}{\sum_{\mathbf{q}_i \in \mathcal{Q}_i} \pi(\mathbf{n})} \quad (17)$$

Note that with (15) inserted, $P_{\mathcal{Q}_i}(\mathbf{q}_i)$ does not contain the normalization constant G which is difficult to determine.

V. MEASUREMENT ERROR AND SIMILAR FLOWS

Now return to the system controlled by MBAC, where due to measurement errors, flows will be accepted also when the system is in the rejection region (Fig. 2) and drive the system above uc . This will again put all the flows at risk of QoS violations. When the QoS requirement is violated the network provides little or no utility to the end user and the network resources can be considered wasted. From a flow point of view the probability of false acceptance should be kept low and this is the focus of this analysis.

The critical situation arises as soon as the system transits from the acceptance region into the rejection region as this is where a false acceptance is first made. In this analysis the state space above uc is omitted. Needless to say, if the probability of false acceptance is unacceptable at the boundary of uc , it is also unacceptable when the system resides above uc .

We shall use the stochastic knapsack defined in the previous section to approximately model the state space within the rejection region. The assumption is then that the impact of the measurement error is not significant when determining the conditional blocking probabilities within this region.

Consider a flow from class i , arriving to this system when the system is in one particular state in the rejection region. Define $P_{FAcc|\mathbf{q}_i}$ as the probability of false acceptance given that the system is in the rejection state $\mathbf{q}_i \in \mathcal{Q}_i$. Let this probability be bounded by the *performance target*, ε_i and define the *conditional performance requirement*:

$$\begin{aligned} P_{FAcc|\mathbf{q}_i} &= P(\text{False acceptance} | \mathbf{N} = \mathbf{q}_i, \mathbf{q}_i \in \mathcal{Q}_i) \\ &= P(\hat{R} + b_i \xi \leq l_{max} \xi | \mathbf{q}_i) \leq \varepsilon_i \end{aligned} \quad (18)$$

$P_{FAcc|\mathbf{q}_i}$ increases as the measurement window size decreases. Because the window size in general is very limited, it may be impossible to meet the above performance target. To cope with this, for a class i flow a safeguard of size $l_i\xi$ is added to make up for the measurement error. Viewing each level as a system resource, where l_{max} is the reserved number of resources to the flows, this implies that a flow from class i will see l_i resources as *unavailable resources*. With an added safeguard l_i a new flow belonging to class i , will only be admitted if

$$\hat{R} + b_i\xi \leq (l_{max} - l_i)\xi, \quad l_i = 0, 1, \dots, l_{max} \quad (19)$$

Including the safeguard, the conditional performance requirement is rewritten:

$$P(\hat{R} + b_i\xi \leq (l_{max} - l_i)\xi \mid \mathbf{N} = \mathbf{q}_i) \leq \varepsilon_i \quad (20)$$

Assume now that the flows are similar according to the definition in Section II. This implies that they all share the auto-correlation function $\Psi(t) = cov(K_i(t), K_i(t + \tau))/\sigma_i^2$. Without the use of similar flows the covariance for every flow must first be determined to find $\zeta_i^2(w)$ separately for each class. Using the property of similar flows significantly simplifies the determination of the variance of the time average of the aggregate rate, which now is directly found by applying (10):

$$\zeta_{\mathbf{n}}^2(w) = \sum_{i=1}^k n_i \sigma_i^2 \frac{2}{w * 2} \int_0^w (w-t)\Psi(t)dt \quad (21)$$

With the assumption from section III-B, that $\hat{R} \sim \mathcal{N}(\xi_{\mathbf{n}}, \zeta_{\mathbf{n}}^2(w))$:

$$\left(\frac{\hat{R} - \xi_{\mathbf{n}}}{\zeta_{\mathbf{n}}(w)} \leq z_{\varepsilon_i} \right) = 1 - \varepsilon_i \quad (22)$$

Rearranging and using the symmetrical properties of the normal distribution:

$$P(\hat{R} \leq \xi_{\mathbf{n}} - \zeta_{\mathbf{n}}(w)z_{\varepsilon_i}) = \varepsilon_i \quad (23)$$

Comparing (23) and (20), the performance target will be met if l_i and $\zeta_{\mathbf{n}}(w)$ satisfy:

$$\xi(l_i + b_i - l_{max}) + \xi_{\mathbf{n}} = \zeta_{\mathbf{n}}(w)z_{\varepsilon_i} \quad (24)$$

With a predefined confidence interval and a fixed window size of w , the required value for l_i given this flow mix is:

$$l_i + b_i = \left\lceil \frac{\zeta_{\mathbf{n}}(w)z_{\varepsilon_i}}{\xi} \right\rceil + b_r \quad (25)$$

where $b_r = l_{max} - \frac{\xi_{\mathbf{n}}}{\xi}$ is the number of levels spanning the rejection region, $0 \leq b_r < b_i$.

Given that the system is in a particular rejection state \mathbf{q}_i , formula (25) can be used to determine the minimum l_i which meets the performance requirement $P_{FAcc|\mathbf{q}_i} \leq \varepsilon$.

Paper [1] gives a detailed analysis of false acceptance when flows are homogeneous. In the homogeneous case, the rejection region only consists of one state $n = l_{max}$.

For this non-homogeneous case, the analysis is more complex since an arriving flow may have several rejection states, where the probability of erroneous decisions depends on the flow mix of the currently accepted flows.

In the real system, MBAC has no other information regarding the system state than the measurement, thus l_i must be valid for any rejection state. Relevant provisioning methods are:

- **Approximate provisioning:** The safeguard for class i , is the smallest l_i which meets the performance requirement:

$$\begin{aligned} P_{F|\mathcal{Q}_i} &= P(\text{False acceptance} \mid \mathcal{Q}_i) \\ &= \sum_{\mathbf{q}_i \in \mathcal{Q}_i} P_{F|\mathbf{q}_i} P_{\mathcal{Q}_i}(\mathbf{q}_i) \leq \varepsilon_i \end{aligned} \quad (26)$$

- **Approximate critical state provisioning:** The safeguard for class i is based on the state \mathbf{q}_i which over the long term results in the highest number of false acceptances:

$$\arg \max_{\mathbf{q}_i \in \mathcal{Q}_i} \{P(\mathbf{N} = \mathbf{q}_i) P_{FAcc|\mathbf{q}_i}\} \quad (27)$$

- **Largest safeguard provisioning:** The safeguard for class i , is based on the rejection state which requires the highest value of l_i :

$$\arg \max_{\mathbf{q}_i \in \mathcal{Q}_i} \left\{ \left\lceil \frac{\zeta_{\mathbf{n}}(w)z_{\varepsilon_i}}{\xi} \right\rceil + b_r \right\} \quad (28)$$

- **Largest variance state provisioning:** The safeguard for class i , is based on the state within the rejection region resulting in the largest variance of the time average of the aggregate mean:

$$\arg \max_{\mathbf{q}_i \in \mathcal{Q}_i} \zeta_{\mathbf{n}} \quad (29)$$

In the case where all states in the rejection region have approximately the same mean rate, provisioning using largest safeguard provisioning and largest variance state is the same. Otherwise, largest safeguard provisioning will be the most pessimistic provisioning method. However, in the case where the state probabilities are not known, this may be the safest method for determining l_i .

Section VI will give a demonstration of the above provisioning methods.

A. MBAC With No Knowledge Regarding Flow Class

We have assumed that the MBAC knows which class a flow belongs to upon flow arrival. It may be that the MBAC cannot distinguish between classes of flows. Flows will then be treated by the MBAC as belonging to the same class, thus the chosen size of slack bandwidth must be common for all In addition, the MBAC is typically fed with the peak rate of the

arriving flow instead of the mean rate. Assuming peak rate, r_i of the arriving flow, adds a pessimism to the MBAC which can be translated to a slack bandwidth of $r_i - \xi_i$. This slack in bandwidth can then be converted to levels.

VI. SIMILAR FLOWS GENERATED BY MMRP SOURCES

The two-state Markov modulated bit-rate process (MMRP) is a simple, yet realistic source model used to model both speech sources and video sources [15]. This is a rate process $K(t) = rI(t)$ where $I(t)$ alternates between states $I = 0$ and $I = 1$ and r is the peak rate. The duration in each state, $I = 0$ and $I = 1$ follows a negative exponential distribution with mean $\frac{1}{\beta}$ and $\frac{1}{\alpha}$ respectively. The auto-covariance of this MMRP process is given by:

$$\text{cov}(K(t), K(t + \tau)) = \sigma^2 e^{-\tau(\alpha + \beta)}. \quad (30)$$

where σ_i^2 is:

$$\sigma_i^2 = \frac{r_i^2 \alpha_i \beta_i}{(\alpha_i + \beta_i)^2} = \frac{r_i^2 \alpha_i (h - \alpha_i)}{h^2} \quad (31)$$

Consider now k classes of similar flows where a flow from class i , is described by the rate process $K_i(t)$. In accordance with the definitions of similar flows, all flows belonging to the same similar flow class, have the same auto correlation function $\Psi(t)$ given by:

$$\Psi(t) = \frac{\text{cov}(K_i(t), K_i(t + \tau))}{\sigma_i^2}, \quad \forall i \quad (32)$$

Let now $K_i(t)$ be generated by a two state MMRP process. If $\beta_i = h - \alpha_i$, see Fig. 3 all flows will have the same $\Psi(t) = e^{-ht}$ and the flows can be classified as similar. Flows belonging to class i is distinguished by the parameters α_i and r_i .

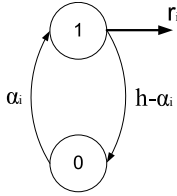


Fig. 3. MMRP source model

The size of the measurement error is expressed through the variance of the estimated mean rate. If the state vector \mathbf{n} is known, the variance of the time average is given by (21):

$$\begin{aligned} \zeta_{\mathbf{n}}^2(w) &= \sum_{i=1}^k n_i \zeta_i^2(w) \\ &= \frac{2}{w^2 h^3} \left(w - \frac{(1 - e^{-wh})}{h} \right) \sum_{i=1}^k n_i r_i^2 \alpha_i (h - \alpha_i) \end{aligned} \quad (33)$$

A. Case study using the two-state MMRP Source Model

In this section we will demonstrate the provisioning methods defined in Section V with a simple example where the flows are similar. Let there be two classes i , $i = 1, 2$ of flows representing real-time video applications (video 1 and video 2) competing for a link controlled by MBAC which admits a flow according to (19). The flows are generated by MMRP processes with parameters shown in Fig. 4, which results in $\xi = 1$ Mbps, $b_1 = 1$ and $b_2 = 5$. According to the definition, the flows can be classified as similar flows. The maximum allowable average rate on this link is $\xi l_{max} = 25$ Mbps, and the estimate of the average aggregate rate is based on continuous observation over a window size of, $w = 10$ s. The task is to control the probability of false acceptance given that the system is in the rejection region, $P_{F|Q_i} < \varepsilon_i$. In this example $\varepsilon_i = 0.025$ for both classes.

Fig. 5, shows the state diagram for this system, where a given state is specified by (n_1, n_2) . Flows representing the video 1 class, have a rejection region of 6 states, corresponding to the states above the solid line:

$$Q_1 = \{(0, 5), (5, 4), (10, 3), (15, 2), (20, 1), (25, 0)\}$$

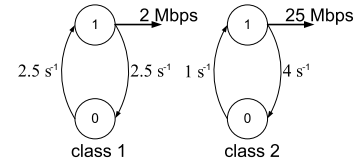


Fig. 4. Parameter settings for class 1 and class 2

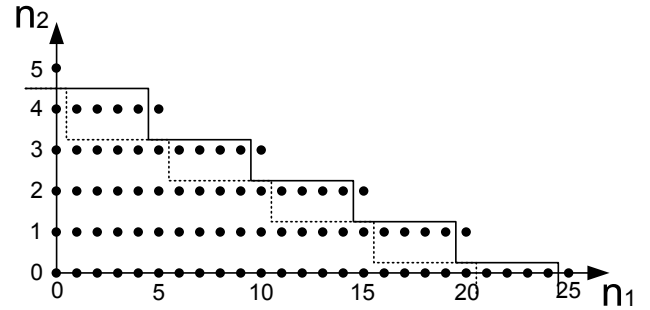


Fig. 5. The rejection region for class 1 are the states above the solid line. The rejection region for class 2 are the states above the broken line

The video 2 class, has a rejection region of 26 states, corresponding to the states above the broken line in Fig. 5.

Let the offered flow load from the video 1 sources and video 2 sources be $A_1 = 20$ erlang and $A_2 = 5$ erlang, respectively. Fig. 6 shows the probability of being in the different rejection states of class 1 conditioned on the system being in the rejection region.

For the video 2 class the conditional probability distribution is presented in Fig. 7-

TABLE I
REQUIRED SAFEGUARD l_i USING DIFFERENT PROVISIONING METHODS WITH THE CORRESPONDING $P(\text{FALSE ACCEPTANCE} \mid \mathcal{Q}_i)$.

Provisioning method	Class 1, (video 1)	Class 2, (video 2)
Approximate provisioning	$l_1 = 5$, $P_{F \mathcal{Q}_1} = 0.015$	$l_2 = 3$, $P_{F \mathcal{Q}_2} = 0.012$
Approximate critical state provisioning	$l_1 = 5$, $P_{F \mathcal{Q}_1} = 0.015$	$l_2 = 4$, $P_{F \mathcal{Q}_2} = 0.0048$
largest safeguard provisioning	$l_1 = 8$, $P_{F \mathcal{Q}_1} = 9.9E^{-4}$	$l_2 = 7$, $P_{F \mathcal{Q}_2} = 2.1E^{-4}$
Largest variance state provisioning	$l_1 = 8$, $P_{F \mathcal{Q}_1} = 9.9E^{-4}$	$l_2 = 4$, $P_{F \mathcal{Q}_2} = 0.0048$

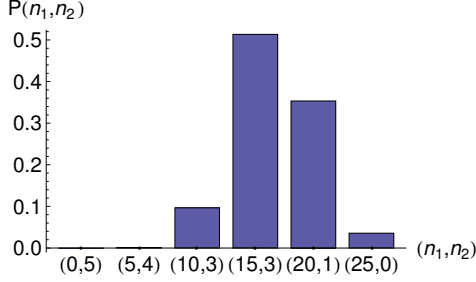


Fig. 6. Probability of being in the different rejection states conditioned on the system being in the rejection region for class 1

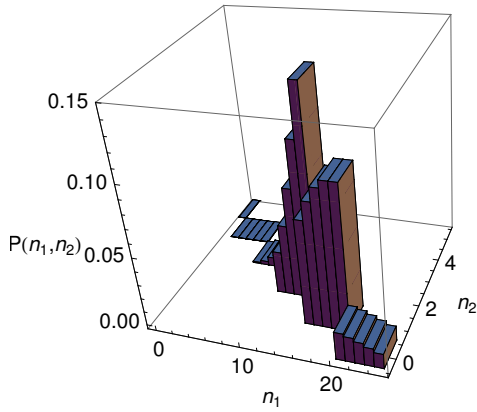


Fig. 7. Probability of being in the different rejection states conditioned on the system being in the rejection region for class 2

Table I shows the required safeguard and the resulting $P_{F|\mathcal{Q}_i}$ for each of the classes using the different provisioning methods defined in Section V.

Consider first provisioning for the video 1 class. If the system only consisted of video 1 sources, only one reduction level would be required in order to meet the performance target, $P_{F|\mathcal{Q}_1} < 0.025$. However, the additional video 2 flows add significant amount of uncertainty to the acceptance decision.

For the video 1 class, since all rejection states have the same mean, the rejection state with the largest variance will also be the rejection state which requires the highest value of l_1 . This will be the state consisting of solely video 2 flows, state (0,5). For largest safeguard provisioning which requires no knowledge of state probabilities, Table ?? shows that 8 levels are required corresponding to a 32% drop in system utilization.

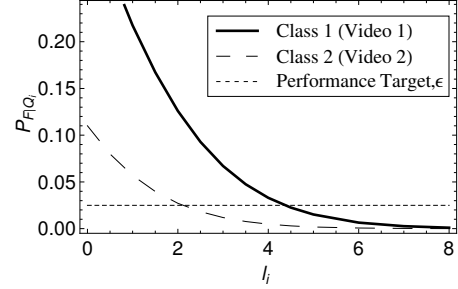


Fig. 8. For class 1 and class 2: the probability of false acceptance given that the system is in the class dependent rejection region $P_{F|\mathcal{Q}_i}$ for different values of the safeguard l_i

System utilization is improved with some knowledge of the state probabilities. The state resulting in the highest number of false acceptances is in this case the most probable state, state (15, 2). Approximate critical state provisioning will thus require a safeguard of size $l_1 = 5$. In this case, the method of approximate provisioning (26) will result in the same $l_1 = 5$ as approximate critical state provisioning.

Fig. 8 shows how the conditional probability of false acceptance, $P_{F|\mathcal{R}_1}$, is reduced as the safeguard increases for both classes. From Fig. 8, it can be seen that $l_1 > 4$ to meet the requirement. Using the approximate or approximate critical state provisioning will both meet the performance target and improve utilization.

Now, move to the video 2 class. For video 2, $b_2 = 5b_1$ and one can think of this as a "pessimism" associated with b_2 corresponding to 5 levels of reduction, 4 more levels than video 1 sources. Due to this effect, as can be seen in Table ??, only a safeguard of size $l_2 = 4$ is required when provisioning based on the largest variance state (0,5). On the other hand now b_r in (25) is no longer always zero.

For video 2, the state which requires the highest value of l_2 , is state (1, 4) and Table ??, shows that largest safeguard provisioning requires $l_2 = 7$ levels. Approximate critical state provisioning requires $l_2 = 4$. A further improvement in terms of utilization can be achieved if approximate provisioning is used. Table ??, shows that the required number of levels is $l_1 = 3$ According to Fig. 8, for the video 2 class, to meet the performance requirement, $l_2 > 2$. Using approximate state provisioning will both meet the performance target and improve utilization.

In this example, the value $P_{F|\mathcal{Q}_i} < 0.025$ was given and the sole purpose was to control the probability of false acceptance

with the condition that the system was in the class dependent rejection region. Whether this is a proper value can only be determined if the complete state space is considered, not just the rejection region at or below uc . The distribution of accepted flows will depend on the flow load A_i from the different classes together with the size of the measurement error. Regardless, 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 [7].

For a given flow load, there will be a safeguard size which balances false acceptances and false rejections. If the safeguard is too large, resources are wasted because flows are rejected unnecessarily. If the safeguard is too small, false acceptances will drive the system above uc , implying QoS violations to the flows and the carried traffic can be considered useless, thus resources are also wasted. The impact of flow dynamics and measurement error on the performance of MBAC in terms of both false acceptance and false rejection is pursued in a separate work.

Simulation is used to check the defined analytical expression (25). For class 2 flows, let the system be in state (0,5). The simulation follows closely what is theoretically predicted when the window size is above 2, see Fig. 9. When the window size is small, correlation between consecutive windows increase the probability of false acceptance.

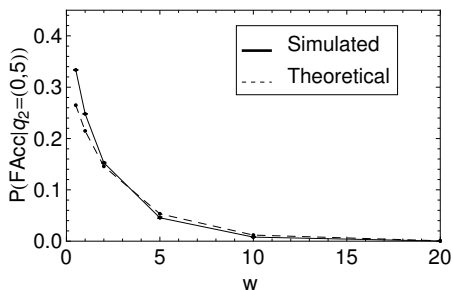


Fig. 9. Theoretically predicted probability of false acceptance for class 2, given that the system is in rejection state (0,5), compared with simulation

VII. CONCLUSION

This current work sheds light on a more in-depth understanding of how measurement errors impact the MBAC acceptance decision. Measurement errors will cause flows to be falsely accepted and falsely rejected. 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 acceptance decision.

In this analysis, the focus was on the probability of false acceptance. Most critical are the system states, within the *rejection region*, where accepting a flow will drive the system to a level beyond its limits. The system can then no longer guarantee QoS to the flows and the service provided to the users become inferior.

By conditioning on being in this rejection region, the task to limit the probability of false acceptance by adding a slack in

bandwidth. With a given probability and window size, the size of this slack can be stated up front for analytically tractable sources with a known covariance function.

By introducing the concept of *similar flows*, the error analysis with non-homogeneous flows is simplified substantially. Similar flows share a common correlation structure and the error analysis becomes straight forward. It can be argued that the assumption of a common correlation structure is a gross simplification, however it is one step ahead of the even more unrealistic assumption that flows are homogeneous.

In order to determine a proper value for the slack bandwidth, the impact of the measurement error on the distribution of accepted flows must be taken into account. If the slack is too large, the probability of false rejections increases and the system utilization decreases. If the slack is too small, the probability of false acceptance impacts the state transitions such that in reality also the state space above the system limits may be visited.

There is a tradeoff between rejecting too many flows thus wasting resources, and accepting too many flows resulting in QoS violations and non-usable carried traffic. A separate work [2] considers this in more detail where the impact of flow arrivals and departures are included in the analysis.

REFERENCES

- [1] A. Nevin, P. J. Emstad, Y. Jiang, and G. Hu, "Quantifying the uncertainty in measurements for mbac," in *Proc. EUNICE 2009*, 2009.
- [2] A. Nevin, P. J. Emstad, and Y. Jiang, "Mbac: Impact of the measurement error on key performance issues," in *Proc. EUNICE 2010*, 2010.
- [3] L. Breslau, S. Jamin, and S. Shenker, "Comments on the performance of measurement-based admission control algorithms," in *IEEE INFOCOM*, 2000.
- [4] A. W. Moore, "Measurement-based management of network resources," Ph.D. dissertation, University of Cambridge, 2002.
- [5] M. Grossglauser and D. N. C. Tse, "A time-scale decomposition approach to measurement-based admission control," *IEEE/ACM Trans. Networking*, vol. 11, no. 4, pp. 550–563, Aug. 2003.
- [6] 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.
- [7] J. W. Roberts, "Internet traffic, QoS and pricing," in *Proceedings of the IEEE*, vol. 92, no. 9, 2004.
- [8] Y. Jiang, A. Nevin, and P. J. Emstad, "Implicit admission control for a differentiated services network," in *Next Generation Internet Design and Engineering, 2006. NGI '06. 2006 2nd Conference on*, 0-0 2006, pp. 8 pp. –365.
- [9] N. G. Bean, "Statistical multiplexing in broadband communication networks," Ph.D. dissertation, University of Cambridge, 1993.
- [10] F. Bricchet, M. Mandjes, and M. F. Sanchez-Canabate, "Admission control in multiservice networks," COST 257, Mid-term Seminar, Interim Report, December 1998.
- [11] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 2nd ed. Springer, 2002.
- [12] R. van de Meent, M. Mandjes, and A. Pras, "Gaussian traffic everywhere?" in *Communications, 2006. ICC '06. IEEE International Conference on*, vol. 2, June 2006, pp. 573–578.
- [13] V. B. Iversen, "Teletraffic engineering and network planning," May 2006, course Textbook.
- [14] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*, P. J. Hancock, Ed. Springer-Verlag London, 1995.
- [15] M. Schwartz, *Broadband integrated networks*, P. Becker, Ed. Prentice Hall, 1996.