Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

Transient Analysis of Traffic generated by Bursty Sources, and its Application to Measurement-based Admission Control

M.R.H. Mandjes, M.J.G. van Uitert

Probability, Networks and Algorithms (PNA)

PNA-R0002 May 31, 2000

Report PNA-R0002
ISSN 1386-3711

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

Transient Analysis of Traffic generated by Bursty Sources, and its Application to Measurement-based Admission Control

Michel Mandjes*, Miranda van Uitert†

*Bell Laboratories, Lucent Technologies
P.O. Box 636, Murray Hill, NJ 07974, USA

†CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

ABSTRACT

The first part of the paper is devoted to a transient analysis of traffic generated by bursty sources. These sources are governed by a modulating process, whose state determines the traffic rate at which the source transmits. The class of modulating processes contains e.g. on/off traffic sources with general on and off times (but is considerably broader). We focus on the probability of extreme fluctuations of the resulting traffic rate, or more precisely, we determine the probability of the number of sources being in the on state reaching a certain threshold, given a measurement of the number of sources in the on state t units of time ago. In particular, we derive large deviations asymptotics of this probability when the number of sources is large. These asymptotics are numerically manageable, and it is empirically verified that they lead to an overestimation of the probability of our interest. The analysis is extended to alternative measurement procedures. These procedures allow to take into account for instance more historic measurements than just one, possibly combined with an exponential weighting of these measurements.

In the second part of the paper, we apply the asymptotic calculation methods to gain insight into the feasibility of measurement-based admission control (MBAC) algorithms for ATM or IP networks. These algorithms attempt to regulate the network's load (to provide the customers with a sufficient Quality of Service), and at the same time achieve an acceptable utilization of the resources. An MBAC algorithm may base acceptance or rejection of a new request on the measured momentary load imposed on the switch or router; if this load is below a given threshold, the source can be admitted. We investigate whether such a scheme is robust under the possible stochastic properties of the traffic offered. Both the burst level (i.e., the distribution of the on and off times of the sources) and the call level (particularly the distribution of the call duration) are taken into account. Special attention is paid to the influence of the bursts, silences, or call durations having a distribution with a 'heavy tail'.

2000 Mathematics Subject Classification: 60F10, 60K25, 90B18.

Keywords and Phrases: transient probabilities, large deviations, ATM and IP networks, measurement-based admission control, call and burst level.

Note: Work of the second author carried out in part under the project PNA2.1 "Communication and Computer Networks". Part of the work was done while both authors were affiliated with KPN Research, P.O. Box 421, 2260 AK Leidschendam, The Netherlands.

1 Introduction

The success of emerging high-speed packet-switched networks strongly depends on the admission control (AC) function. This function decides on whether or not to accept a newly arrived traffic flow (also referred to as *call*). A request should be rejected if it would lead to violation of the quality of service (QoS) guarantees of either the existing flows or the flow itself. On the other hand, the AC function must be designed such that bandwidth utilization is sufficiently high. Notice that these two objectives are fundamentally conflicting: in order for an AC function to guarantee a satisfactory level of QoS it would be tempting to be conservative in accepting flows, whereas utilization maximization would ask for a more aggressive acceptance policy.
In ATM (Asynchronous Transfer Mode) networks connections are established, and therefore we can speak of *Connection Admission Control*. CAC has always played a crucial role in ATM traffic management, see for instance Chapter 5 of [25]. In IP (Internet Protocol) based networks, recently QoS classes were defined, giving rise to the use of admission control. In particular, within the Intserv working group of the Internet Engineering Task Force, the so-called Controlled Load service class is being developed [28]. Although this class does not strictly require a specific target QoS performance (like in ATM), operationally the success of the service is determined by the QoS offered, justifying the use of an admission control function.

Rationale for measurement-based AC. It is not that difficult to design a 'safe' AC function (i.e., an AC function delivering sufficient QoS): each flow is simply allocated its peak rate, and a flow is accepted as long as the sum of the peak rates is below the link capacity. In general however, traffic flows do not use that peak rate all the time: they typically tend to offer *bursty* traffic patterns, to be modeled by, e.g., on/off flows. For that reason, traffic descriptors have been developed to deal with these variable bit rate streams. The idea is to describe the traffic not only by its peak rate, but also by a sustained rate (which is an upper bound to the long-run average), and a maximum burst size (which is the largest amount of traffic that can consecutively be sent at peak rate). The compliance of the traffic flow to these parameters can be enforced by policing. The worst-case traffic fitting into the descriptor can be determined, and an admission control function based on this profile will still certainly be safe [13]. The drawback, however, is that measurements show that the three-parameter traffic descriptor is not 'tight' at all, as it is a loose envelope of the real traffic pattern. In other words: the traffic actually sent is in general considerably 'better' than this worst-case traffic. Consequently, the resulting AC function cannot be efficient.
A novel idea to exploit the gap between the worst-case profile and the real traffic pattern, is to do admission control based on real-time traffic measurements [7, 14, 17, 18]. Its objective is to increase bandwidth utilization while maintaining the desired QoS level. It should be emphasized that such a *measurement-based admission control* (MBAC) function cannot *strictly guarantee* QoS, as it is only based on empirical information. For that reason, such an MBAC is not to be used for flows that cannot afford any violation of their QoS. However, the efficiency gain can be large enough that it is preferred to traditional, worst-case profile based AC functions.

Scope of the paper. The crucial feature examined in the present paper is the robustness of MBAC algorithms with respect to the characteristics of the traffic patterns offered. To explain

the issues involved, we refer to an early paper on MBAC by Gibbens, Kelly, and Key [15]. They claim that the algorithm they propose does not rely critically upon traffic assumptions. Their line of reasoning is the following.
Calls arrive according to a Poisson process at a switch or router and holding times are i.i.d. samples from a general distribution. The resource is assumed to be bufferless, meaning that traffic is lost if the aggregate input rate exceeds the link capacity C. Calls are mutually independent and statistically identical. They are of the on/off type, with known peak rate.
Suppose the mean rate can be captured by measurements. Taking also into account that no buffer capacity is available, it is well known that the loss probability for a fixed number of calls does not depend on the entire distribution of the on and off times, but only on the mean and peak rate. Its stationary distribution only depends on the call holding time distribution through its mean. The resulting *insensitivity* to both the on and off times, and the call holding time would make their admission schemes not critically dependent upon traffic assumptions.
Within this framework, the following MBAC algorithm is proposed by Gibbens, Kelly and Key [15]. Given that n calls are currently in progress, a threshold $s(n)$ is derived: if the current measured load is below this $s(n)$ a new call can be admitted. Several algorithms to find this $s(n)$ are proposed; two of them are based on the (transient) probability that the aggregate input rate exceeds the link rate at some point in time t, given the measured mean of the calls and the load offered to the ATM switch or IP router at time 0. If this probability is below, say, 10^{-6}, the new call can be accepted. We have two major objections against this approach.

- In the first place, *stationary* loss probabilities are independent of the distributions of the on and off-times, but *transient* loss probabilities are *not*. Obviously, a given mean input rate only determines the ratio between the mean on and off times, but does not relate to their absolute values. As a consequence it is clear beforehand that the MBAC algorithm does not work for all on/off inputs with the same mean rate, as was claimed by the authors of [15]. However, the major question we ask ourselves is: how robust is the procedure

  with respect to the shapes of the distributions of the bursts and silences? Therefore, we chose to hold the respective means fixed (so *a fortiori* the same mean rate).

- In the second place, no guideline is given on how to choose the parameter t. Obviously, this parameter should relate to the call level, as future cleared down calls may significantly affect the transient overflow probability. One would expect that the shape of the call duration distribution does have influence here.

It is clear that knowledge of the *transient* of the traffic rate generated by a set of traffic flows provides useful input to the question whether MBACs similar to the one of [15] are feasible. To study this matter, we first develop a method to calculate the loss probability in a time interval following a traffic measurement. Then we examine the sensitivity of this probability to the burst level and call level stochastics.

Mathematical model. In order to analyze the relevant transient probabilities, we consider a model consisting of a number of, say n, flows. Assume that any flow is equipped with a d-dimensional state space, where each state corresponds to a specific constant traffic rate. The consecutive sojourn times in state i ($i = 1, \ldots, d$) have general distribution S_i and are i.i.d. We assume that the flows feed into a queue with a buffer of negligible size and link rate C; consequently, buffer overflow corresponds to an aggregate input rate exceeding C. The most

appealing example consists of flows with state spaces of only two traffic levels: an on and off state; we will explain all concepts on the basis of these on/off flows.
Suppose that at time 0 it is observed that m_0 flows (out of the n flows that have been accepted) are in the on-state. Accepting a new flow (say at time 0) means that the system has to cope with the traffic generated by these n flows *and* the new flow. As the system cannot terminate on-going calls, the only measure it can take after accepting the new request is to block all requests after time 0. Therefore we will calculate the overflow probability, being defined as the probability of the traffic rates of the flows (including the one to be accepted) reaching the link rate at some time $t > 0$, given that all requests arriving after time 0 are rejected. Notice that the calculation of this probability requires knowledge of both burst level and call level dynamics: the termination of calls (and particularly the shape of the call duration distribution) may have a significant impact.
It is intuitively clear that, for a given n, the overflow probability increases with m_0. For an MBAC algorithm it would be attractive to be able to determine a specific value of $m_0 = s(n)$ such that new flows can be admitted as long as this threshold is not reached, and should be rejected otherwise. A number of interesting questions are:

- How sensitive is the value of $s(n)$ to the distribution of the on and off times? This can be examined by calculating the overflow probability for different distributions of on times and off times. We specifically selected the on and off times exponentially distributed and heavy-tailed, in order to investigate their impact. The reason for distinguishing between these particular types of distributions is that exponential tails are known to behave fundamentally different from heavy tails: a superposition of a large number of flows with specific heavy-tailed bursts is known to show long-range dependence, see e.g. [4], whereas exponential bursts do not yield this property.

- How sensitive is the value of $s(n)$ to the holding time distribution of the flows? To answer this question, again the overflow probability can be investigated. Given a distribution of the on and off times, the overflow probability has to be calculated for different distributions of the holding times (again the mean held fixed). We are especially interested in the sensitivity of the value of $s(n)$ with respect to holding times with an exponential and heavy-tailed distribution.

- If there is a strong sensitivity of the value of $s(n)$ to the shape of the distributions, one might wonder what the *best* value is, i.e. a value that is safe for all (realistic) types of traffic. In particular, we wonder how the MBAC algorithm suffers from the consequences of distributions with heavy tails.

Existing literature and contribution. Our analysis of the probabilities of extreme input rate fluctuations is of an asymptotic nature. We derive a large deviations approximation [27] of the transient overflow probability asymptotically in the number of flows. In earlier work [22] we restricted ourselves to the case where the sojourn times in the states of the flow had an exponential distribution; the analysis provided here holds for generally distributed sojourn times. We present results for a variety of measurement procedures. The results are related to those of Duffield [9] who considers the situation with a non-negligible queue, and with Duffield and Whitt [11] who develop procedures to approximate future aggregate input (with techniques related to the ones we use in the present paper, like inversion of Laplace transforms). Where

several previous studies on MBAC were quite heuristic, see e.g. Jamin *et al.* [18], we attempt to provide a more thorough basis.
The goal of the present paper is *not* to provide an MBAC algorithm itself, but rather to shed light on the *feasibility* of such mechanisms: given the intrinsic unpredictability of the input traffic, can a simple threshold policy be a good admission criterion? For simplicity reasons our model only deals with a single QoS class (as opposed to [18] where a multiple buffer structure is considered), that uses a bufferless resource (unlike [14] who consider the case of a non-zero buffer). Our approach is different from [7], who essentially use the 'traditional' effective bandwidth formulae, in which measured means are inserted.
In the present study, we assume that some basic properties of the network traffic are known from historic measurements: reliable values of the mean burst, silence and call duration are available. We try to exploit the dynamics of the system by measuring the current load, cf. the procedures proposed in Mitra, Reiman, and Wang [24]. This focus is fundamentally different from recent work on MBAC algorithms by Duffield [10] and Grossglauser and Tse [16]. They concentrate on the estimation of traffic characteristics out of measurements; the intrinsic unreliability of the resulting estimates is taken into account in their admission procedures.
We believe that the most significant contribution of the paper is the insight into the influence of the shapes of the distributions. Strikingly, as already indicated by Duffield and Whitt [11], heavy tails on the burst level (burst and silences) cause the measurement m_0 to be a relatively good predictor of future load, while a heavy-tailed call duration cannot do much harm as long as there is a good 'separation of time scales': we provide thorough justification why the distribution of the call duration only influences the overflow probability through its mean, as long as the mean call duration is significantly larger than the mean burst and silence.

The organization of this paper is as follows. In Section 2, we first give a model description. Then we derive the asymptotics of the transient overflow probability as defined above, assuming that calls do not leave the system after time 0. Hence, we determine the influence of the on and off period distribution on the transient overflow probability and postpone the influence of the call holding time distribution on this probability to Section 3. We also comment on the extension of the result to more general load measurement procedures, like the ones proposed in [7, 18]. Next to the derivation of an expression for the transient overflow probability taking into account call holding times, Section 3 focuses on the application to MBAC algorithms and provides numerical examples. Section 4 presents the conclusions.

2 Traffic rate fluctuations

In this section we present a method to calculate the transient probability of the input rate of a switch or router attaining an extreme value at time t, given an observed input rate at time 0. We assume in this section that the n flows present do not leave the system. Consequently we will disregard call level effects and we will only examine the impact of the distributions of the bursts and silences. The call level is taken into account in Section 3. Notice that it is practically not feasible to observe an input rate in an infinitesimally small time interval; therefore we also present an analogous procedure in which we condition on the traffic that arrived in an interval. Finally we will describe the methods to numerically capture the transient probabilities.

2.1 Model description

We consider a model consisting of n traffic flows. Any flow is governed by a *modulating process* that can be in d states. In state i traffic is generated at a constant rate r_i. The sojourn time in state i is distributed as a random variable S_i; we assume these sojourn times to have finite means. After completion of the sojourn time, the process jumps to state j with probability q_{ij}. Denote by $\tilde{\pi}$ the invariant distribution of the discrete time Markov chain with transition matrix $Q := (q_{ij})_{i,j=1}^{d}$, i.e., $\tilde{\pi}$ solves $\tilde{\pi} = \tilde{\pi}Q$. Then it is straightforward to see that the equilibrium distribution of the modulating process is

$$\pi_i = \frac{\tilde{\pi}_i E(S_i)}{\sum_{j=1}^{d} \tilde{\pi}_j E(S_j)}.$$

We denote by $p_{ij}(t)$ the transient probability of the modulating process being in state j at time t, given that it was in state i at time 0.

2.2 Transient probability based on an instantaneous measurement

Assume that $A_n(t)$ is a d-dimensional vector denoting the empirical distribution of the n flows, e.g., the jth component of $A_n(t)$ gives the number of flows in state j at time t. We are interested in the probability of having an empirical distribution equal to $n\beta$ at time t, given the empirical distribution $n\alpha$ at time 0:

$$q_n(t, \alpha, \beta) := \mathrm{P}(A_n(t) = n\beta | A_n(0) = n\alpha). \tag{1}$$

This probability decays exponentially in the number of flows. Hence it is essentially determined by its decay rate. Note that the elements of both the vector α and the vector β sum to 1. With the interpretation of y_{ij} being the fraction of flows that goes from state i at time 0 to state j at time t, the following proposition gives the decay rate of probability (1).

Proposition 2.1 *The decay rate is given by*

$$\lim_{n\to\infty} \frac{1}{n} \log P(A_n(t) = n\beta | A_n(0) = n\alpha) = \max_{y_{ij}} \left[\sum_{i=1}^{d} \sum_{j=1}^{d} y_{ij} \log \left(\frac{p_{ij}(t)\alpha_i}{y_{ij}} \right) \right], \tag{2}$$

where $y_{ij} \in [0, 1]$ *is subject to*

$$\sum_{i=1}^{d} y_{ij} = \beta_j,\ j \in \{1, \ldots, d, \}\ \textit{and}\ \sum_{j=1}^{d} y_{ij} = \alpha_i,\ i \in \{1, \ldots, d\}. \tag{3}$$

Proof. We will prove this proposition using 4 steps.
• **Step 1.** In this part of the proof we will derive an exact expression for (1). We will use the following random variables

$$I_{ik} = \begin{cases} 1 & \text{if flow } k \text{ is in state } i \text{ at time 0,} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$J_{ik} = \begin{cases} 1 & \text{if flow } k \text{ is in state } i \text{ at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

Let Y_n be the set of $a_{ik}, b_{ik} \in \{0,1\}$ $(k \in \{1,\ldots,n\}, i \in \{1,\ldots,d\})$, with a_{ik} and b_{ik} the realizations of I_{ik} and J_{ik} respectively, such that $\sum_{k=1}^{n} a_{ik} = n\alpha_i$ and $\sum_{k=1}^{n} b_{ik} = n\beta_i$ for every i. If we take into account the per flow realization probability (1) is equal to

$$\frac{\mathrm{P}(A_n(t) = n\beta, A_n(0) = n\alpha)}{\mathrm{P}(A_n(0) = n\alpha)} = \sum_{a,b \in Y_n} \frac{\prod_{k=1}^{n} \mathrm{P}(I_{ik} = a_{ik}, J_{ik} = b_{ik}, \forall i \in \{1,\ldots,d\})}{\mathrm{P}(A_n(0) = n\alpha)}.$$

Rewriting this expression we obtain

$$\sum_{a,b \in Y_n} \frac{\prod_{k=1}^{n} \mathrm{P}(J_{ik} = b_{ik}, \forall i | I_{ik} = a_{ik}, \forall i) \mathrm{P}(I_{ik} = a_{ik}, \forall i)}{\mathrm{P}(A_n(0) = n\alpha)}. \tag{4}$$

The probability of being in state j at time 0 is equal to the equilibrium probability π_j of being in state j. Therefore $A_n(0)$ follows a multinomial distribution with parameters n and π_1, ..., π_d. Thus (4) equals

$$\sum_{a,b \in Y_n} \frac{\prod_{k=1}^{n} \mathrm{P}(J_{ik} = b_{ik}, \forall i \in \{1,\ldots,d\} | I_{ik} = a_{ik}, \forall i \in \{1,\ldots,d\}) \pi_1^{n\alpha_1} \ldots \pi_d^{n\alpha_d}}{\binom{n}{n\alpha_1,\ldots,n\alpha_d} \pi_1^{n\alpha_1} \ldots \pi_d^{n\alpha_d}}.$$

Before we further evaluate this expression, we define x_{ij} as the number of flows that go from state i at time 0 to state j at time t. Abbreviating $p_{ij} := p_{ij}(t)$, we can rewrite the numerator of (4) as a multinomial probability

$$q_n(t,\alpha,\beta) = \sum_{x \in X_n} \frac{\binom{n}{x_{11},\ldots,x_{dd}} p_{11}^{x_{11}} \ldots p_{dd}^{x_{dd}}}{\binom{n}{n\alpha_1,\ldots,n\alpha_d}}, \tag{5}$$

where the multinomial coefficient counts all possible permutations in which the x_{ij} can be taken out of the n flows. Here X_n is the set of integer x_{ij} such that $\sum_j x_{ij} = n\alpha_i$ and $\sum_i x_{ij} = n\beta_j$.

• **Step 2.** In this part we will derive an asymptotic expression for (5) using a scaling: we will replace x_{ij} by ny_{ij}. Then we obtain for (5)

$$\sum_{ny \in X_n} \frac{(n\alpha_1)! \ldots (n\alpha_d)! p_{11}^{ny_{11}} \ldots p_{dd}^{ny_{dd}}}{(ny_{11})! \ldots (ny_{dd})!}.$$

Recall Stirling's formula for the factorial function $n! = \sqrt{(}2\pi n) n^n e^{-n} e^{\theta(n)}$ with $\theta(n) \in [0, 1/(12n)]$, see formula 6.1.38 in [2]. Then we obtain for (5)

$$q_n(t,\alpha,\beta) = \sum_{ny \in X_n} \frac{e^{S(y)}}{(\sqrt{2\pi})^{(d^2-d)}} \frac{(n\alpha_1)^{n\alpha_1+1/2} \ldots (n\alpha_d)^{n\alpha_d+1/2} p_{11}^{ny_{11}} \ldots p_{dd}^{ny_{dd}}}{(ny_{11})^{ny_{11}+1/2} \ldots (ny_{dd})^{ny_{dd}+1/2}},$$

where $S(y) := \theta(n\alpha_1) + \ldots + \theta(n\alpha_d) - \theta(ny_{11}) - \ldots - \theta(ny_{dd})$.

• **Step 3.** We will show that the decay rate of $q_n(t,\alpha,\beta)$ equals

$$\lim_{n\to\infty} \frac{1}{n} \log \left[\max_{ny \in X_n} \left(\frac{e^{S(y)}}{(\sqrt{2\pi})^{(d^2-d)}} \frac{\alpha_1^{n\alpha_1+1/2} \ldots \alpha_d^{n\alpha_d+1/2} p_{11}^{ny_{11}} \ldots p_{dd}^{ny_{dd}}}{y_{11}^{ny_{11}+1/2} \ldots y_{dd}^{ny_{dd}+1/2}} \right) \right]. \tag{6}$$

That is, we will show that the decay rate of $q_n(t,\alpha,\beta)$ is equal to the decay rate of the largest term of $q_n(t,\alpha,\beta)$, also known as Laplace's principle (see for an extensive treatment of this principle Dupuis and Ellis [12]).

Let us first find an upper bound on the number of y_{ij} such that $ny_{ij} \in X_n$. As we have that ny_{ij} must attain an integer value between 0 and n, such an upper bound is $(n+1)^{d^2}$. This means that an upper bound for the decay rate of $q_n(t, \alpha, \beta)$ is expression (6), increased by

$$\lim_{n\to\infty} \frac{1}{n} d^2 \log[n+1] = 0.$$

Thus, the upper bound equals (6). The lower bound for the decay rate of $q_n(t, \alpha, \beta)$ is trivial: the sum is larger than the maximum.

- **Step 4.** In this step the proof will be completed. Because

$$\lim_{n\to\infty} \frac{1}{2n}(d^2 - d)\log(2\pi) = 0, \quad \lim_{n\to\infty} \frac{1}{n} \max_{ny\in X_n} S(y) = 0,$$

and

$$\lim_{n\to\infty} \frac{1}{2n} \max_{ny\in X_n} \log\left(\frac{\alpha_1 \dots \alpha_d}{y_{11} \dots y_{dd}}\right) = 0,$$

and interchanging log and max, equation (6) equals, y_{ij} satisfying (3):

$$\lim_{n\to\infty} \frac{1}{n} \log q_n(\alpha, \beta) = \max_{y_{ij}} \sum_{i=1}^{d}\sum_{j=1}^{d} y_{ij} \log\left(\frac{p_{ij}\alpha_i}{y_{ij}}\right).$$

□

2.3 Calculation of the decay rate

We found an expression for the decay rate of probability (1) in Section 2.2, which still contains a maximization problem. This problem can easily be solved using the Lagrange function $L(y_{ij}, \lambda_j, \mu_j) :=$

$$\sum_{i=1}^{d}\sum_{j=1}^{d} y_{ij} \log\left(\frac{p_{ij}(t)\alpha_i}{y_{ij}}\right) + \sum_{j=1}^{d} \lambda_j \left(\sum_{i=1}^{d} y_{ij} - \beta_j\right) + \sum_{i=1}^{d} \mu_i \left(\sum_{j=1}^{d} y_{ij} - \alpha_i\right).$$

Differentiating L with respect to all its arguments, and equating the derivatives to zero, gives the optimizing y_{ij} to be equal to $y_{ij} = p_{ij}(t)\alpha_i e^{\lambda_j + \mu_i - 1}$. Taking $\zeta_j := \lambda_j - 1$ we obtain $y_{ij} = \alpha_i e^{\mu_i} p_{ij}(t) e^{\zeta_j}$, where μ_i and ζ_j have to satisfy the following conditions

$$\sum_{i=1}^{d} \alpha_i e^{\mu_i} p_{ij}(t) e^{\zeta_j} = \beta_j,\ j = 1, \dots, d, \quad \text{and} \quad \sum_{j=1}^{d} e^{\mu_i} p_{ij}(t) e^{\zeta_j} = 1,\ i = 1, \dots, d.$$

Using these expressions for y_{ij}, μ_i and ζ_j the decay rate of $q_n(t, \alpha, beta)$ has a nice symmetric form

$$\sum_{i=1}^{d}\sum_{j=1}^{d} \alpha_i e^{\mu_i} p_{ij}(t) e^{\zeta_j} \log(e^{-\mu_i} e^{-\zeta_j}) =$$

$$-\sum_{i=1}^{d} \alpha_i \mu_i \left(\sum_{j=1}^{d} e^{\mu_i} p_{ij}(t) e^{\zeta_j}\right) - \sum_{j=1}^{d} \zeta_j \left(\sum_{i=1}^{d} \alpha_i e^{\mu_i} p_{ij}(t) e^{\zeta_j}\right) = -\sum_{i=1}^{d} (\alpha_i \mu_i + \beta_i \zeta_i).$$

Remark 2.1. Proposition 2.1 is a generalization of the main result in earlier work Mandjes [22], where we examined the special case of the sojourn times $(S_i)_{i=1}^d$ having exponential distributions. For that case, we could also find the 'most probable path' from empirical distribution $n\alpha$ to $n\beta$, taking t units of time. We emphasize that in the present paper, we developed a more direct approach, starting with an exact expression (5), and finding its decay rate without invoking any large deviations theorem.

Remark 2.2. In the case when flows have a two-dimensional state space (where the states are denoted by 1 and 0) we can express y_{01}, y_{10} and y_{11} in terms of y_{00} using (3). Denoting $p := [p_{01}(t)p_{10}(t)]^{-1}p_{00}(t)p_{11}(t)$, the optimizing y_{00} equals

$$\begin{cases} 1-\alpha_1-\beta_1+\alpha_1\beta_1 \\ \qquad \text{if } p=1, \\ -(p(\alpha_1-2+\beta_1)+1-\alpha_1-\beta_1)\,(2(p-1))^{-1}+ \\ \sqrt{p^2(\alpha_1-\beta_1)^2+2p(\alpha_1+\beta_1-\alpha_1^2-\beta_1^2+(1-\alpha_1-\beta_1)^2}\,(2(p-1))^{-1} \\ \qquad \text{otherwise.} \end{cases}$$

2.4 Transient probability based on alternative measurement procedures

In the previous subsections it was explained how to calculate probabilities of extreme traffic rate fluctuations, under the assumption that it is feasible to make a reliable instantaneous measurement of the empirical distribution of the flows at time 0. In practice however, one will measure the traffic that arrived in an interval rather than at just one point in time. Analogously to the previous section, we can derive the decay rate of reaching a certain traffic rate at time t, given a measurement in the interval $[-s,0]$.
As was described above, in this section we analyze the probability

$$\mathrm{P}(A_n(t)=n\alpha|R\{[-s,0]\}=n\kappa), \tag{7}$$

where $R\{[0,t]\}$ is defined as $\sum_{k=1}^n R_k\{[0,t]\}$, and the $R_k\{[0,t]\}$ are functions of the kth modulated renewal process $X_k(s)$, for s in $[0,t]$. Later on we will give a number of examples of $R_k\{[0,t]\}$. The decay rate of (7) can be rewritten as (apply Bayes' rule)

$$\lim_{n\to\infty}\frac{1}{n}\log\mathrm{P}\left(\sum_{k=1}^n R_k\{[-s,0]\}=n\kappa|A_n(t)=n\alpha\right)$$

$$+\lim_{n\to\infty}\frac{1}{n}\log\mathrm{P}\left(A_n(t)=n\alpha\right)-\lim_{n\to\infty}\frac{1}{n}\log\mathrm{P}\left(\sum_{k=1}^n R_k\{[-s,0]\}=n\kappa\right) \tag{8}$$

These three decay rates can be calculated as follows.

- In Section 2.1, we have described the process $((p_{ij})_{i,j=1}^d,(S_i)_{i=1}^d)$. It is not difficult to see that the time-reversal of this process is $((\tilde{p}_{ij})_{i,j=1}^d,(S_i)_{i=1}^d)$, where $\tilde{p}_{ij}:=p_{ji}\pi_j/\pi_i$. Denote by $\tilde{R}_k\{[0,t]\}$ the analogue of $R_k\{[0,t]\}$, but then with the underlying time-reversed process. Consequently, the first decay rate equals

$$\lim_{n\to\infty}\frac{1}{n}\log\mathrm{P}\left(\sum_{k=1}^n \tilde{R}_k\{[0,s]\}=n\kappa|A_n(-t)=n\alpha\right).$$

Using Cramér's theorem [27, p. 14], cf. equation (26) of [13], this reads

$$\sup_{\theta}\left(\theta\kappa - \sum_{i=1}^{d}\alpha_i \log \mathrm{E}\left(\exp[\theta\tilde{R}_k\{[0,s]\}]|X_k(-t)=i\right)\right).$$

- The second decay rate is simply an application of Sanov's theorem [27, p. 37]: the discrepancy between measure α and invariant π is given by decay rate $\sum_{i=1}^{d}\alpha_i \log(\alpha_i/\pi_i)$.
- The third decay rate is again found using Cramér's theorem:

$$\sup_{\theta}\left(\theta\kappa - \log\left(\sum_{i=1}^{d}\pi_i \mathrm{E}\left(\exp[\theta R_k\{[0,s]\}]|X_k(-t)=i\right)\right)\right).$$

Examples of measurement procedures. Recall that $R_k\{[0,t]\}$ is a function of the kth modulated renewal process $X_k(s)$ for s in $[0,t]$, where $X_k(s)$ is the state of call k at time s. For the sake of convenience we confine ourselves to the modulated process belonging to call 1, $X_1(s)$. It should be emphasized that several measurement schemes – as proposed in the literature – fit into the framework described above. Among these measurement schemes are the following possible choices of $R_1\{[0,t]\}$.

- The most fundamental choice is the amount of traffic generated in $[0,t]$

$$R_1\{[0,t]\} := \int_0^t r_{X_1(s)}\mathrm{d}s.$$

- Brichet and Simonian [7] propose an exponentially weighted moving average scheme

$$R_1\{[0,t]\} := \alpha\int_0^t r_{X_1(s)}\mathrm{d}s + (1-\alpha)R_1\{[-t,0]\},$$

  for $\alpha \in [0,1]$. If $\alpha > \frac{1}{2}$ this algorithm gives more weight to the recent history than to older measurements.

- Jamin *et al.* [18] propose to improve on an interval measurement, in order to make it more conservative (and to avoid acceptance of flows without having sufficient resources available). They do so by splitting the interval of length t into N subintervals of length $s := t/N$, and taking the maximum rate on these intervals

$$R_1\{[0,t]\} := \max\left(\int_0^s r_{X_1(u)}\mathrm{d}u, \cdots, \int_{t-s}^t r_{X_1(u)}\mathrm{d}u\right).$$

2.5 Numerical methods

In order to numerically evaluate the decay rates derived in the previous subsections, we need to evaluate quantities such as $p_{ij}(t)$ and the moment generating function of the amount of traffic generated in an interval of given length. In fact, there are two situations that allow for explicit analysis, namely (i) the situation of general dimension d and exponential sojourn times in the on and off state, and (ii) the one with two states and generally distributed on and off times.

General d, exponential sojourn times. Denoting the generator of the underlying continuous time Markov chain by $\Lambda := (\lambda_{ij})_{i,j=1}^{d}$, it is well known that $p_{ij}(t) = (\exp(\Lambda t))_{ij}$. With this expression, we can perform the calculation of the transient probability based on the instantaneous measurement. For the alternative measurement procedures, we rely on the following formula, see e.g. [6, 19],

$$\mathrm{E}\left(\exp\left[\theta \int_0^t r_{X_k(s)} \mathrm{d}s\right] |X_k(0) = i\right) = \sum_{j=1}^{d} \exp((\Lambda + \theta R)t)_{ij}.$$

Two states, general sojourn times. We assume that the flows can only be in an on state or off state, denoted by 1 and 0 respectively. Consequently we need the probabilities $p_{ij}(t)$, $i, j \in \{0, 1\}$, to determine probability (1). Here we confine ourselves to $p_{00}(t)$, since this also yields $p_{11}(t)$ by interchanging the roles of the on and off times, and $p_{01}(t) = 1 - p_{00}(t)$ and $p_{10}(t) = 1 - p_{11}(t)$.
We first give a derivation of the Laplace transform (LT) of probability $p_{00}(t)$ analogously to Section 2.1 of Cohen [8]. First, we determine the 'simultaneous' LT of time t and the amount of time in the on state during $[0, t]$, o_t

$$\int_0^\infty e^{-st} \mathrm{E}\left[e^{-\rho o_t} I_{\{X_k(t)=0\}} | X_k(0) = 0\right] \mathrm{d}t. \tag{9}$$

Using this expression, we can determine a formula for the LT of the transient probability $p_{00}(t)$. We finally describe the numerical inversion of this formula, that enables us to capture the probability $p_{00}(t)$.
Let $\hat{S}_i$ be the random variable denoting the residual sojourn time in state i, which is known to have density

$$f_{\hat{S}_i}(x) := \frac{1}{E[S_i]}(1 - F_{S_i}(x)),$$

where F_{S_i} is the probability distribution function of S_i. Denoting by $S_{j,l}$ the lth sojourn time of a call in state j, the simultaneous moment generating function of o_t and time t is given by

$$\mathrm{E}\left[e^{-\rho o_t} I_{\{X_k(t)=0\}} | X_k(0) = 0\right] = \mathrm{P}\left(\hat{S}_0 \geq t\right) +$$

$$\sum_{n=1}^{\infty} \int_{0^-}^{t} e^{-\rho\nu} \mathrm{P}\left(\sum_{l=1}^{n-1} S_{0,l} + \hat{S}_0 \leq t - \nu, \sum_{l=1}^{n} S_{0,l} + \hat{S}_0 \geq t - \nu\right) \mathrm{dP}\left(\sum_{l=1}^{n} S_{1,l} \leq \nu\right). \tag{10}$$

Let F_X^* denote the LT of the probability distribution of X and f_X^* the LT of the density of X. Using straightforward calculation we obtain the LT of the second term of (10)

$$\int_0^\infty e^{-st} \mathrm{P}\left(\hat{S}_0 \geq t\right) \mathrm{d}t = \frac{1}{s} - \frac{1 - f_{S_0}^*(s)}{\mathrm{E}[S_0]s^2}.$$

Interchanging the order of integration, the LT of the first term of (10) yields

$$\sum_{n=1}^{\infty} \left(F^*_{\sum_{l=1}^{n-1} S_{0,l} + \hat{S}_0}(s) - F^*_{\sum_{l=1}^{n} S_{0,l} + \hat{S}_0}(s)\right) \int_0^\infty e^{-(s+\rho)\nu} \mathrm{dP}\left(\sum_{l=1}^{n} S_{1,l} \leq \nu\right),$$

which is equal to

$$\sum_{n=1}^{\infty} \frac{1}{s} \left((f_{S_0}^*(s))^{(n-1)} - (f_{S_0}^*(s))^n \right) \frac{1 - f_{S_0}^*(s)}{\mathrm{E}[S_0]s} (f_{S_1}^*(s+\rho))^n.$$

After some algebra we obtain for (9)

$$\frac{1}{s} - \frac{(1 - f_{S_1}^*(s+\rho))(1 - f_{S_0}^*(s))}{\mathrm{E}[S_0]s^2(1 - f_{S_0}^*(s) f_{S_1}^*(s+\rho))}.$$

We see that this LT is expressed completely in terms of the LTs of the on and off densities. Inserting $\rho = 0$ in (9) gives the LT $\int \exp[-st] p_{00}(t) \mathrm{d}t$. The final step is then to use the numerical inversion formula of the Laplace transform given in Abate and Whitt [1], formula (5.27)

$$f(t) = \frac{e^{\frac{a}{2}}}{2t} f^* \left(\frac{a}{2t} \right) + \frac{e^{\frac{a}{2}}}{t} \sum_{r=1}^{\infty} (-1)^r \mathrm{Re} \left(f^* \left(\frac{a + 2r\pi\sqrt{-1}}{2t} \right) \right) - \epsilon(a,t),$$

for $t > 0$, any constant $a > 0$ and error term $\epsilon(a,t)$,

which gives the transient probability $p_{00}(t)$. Too small an a, causes too large errors, as argued by Abate and Whitt [1]. On the other hand a cannot be too large because of computational difficulties. Since it is difficult to determine the infinite series $\sum_{r=0}^{\infty} (-1)^r w_r$ numerically, we have used the Euler Sum $\mathrm{E}(m,n)$ as approximation

$$\mathrm{E}(m,n) = \sum_{k=0}^{m} \binom{m}{k} 2^{-m} \sum_{g=0}^{n+k} (-1)^g w_g,$$

with appropriate choices for m and n.

3 Application to MBAC

In this section we again consider the decay rate of the overflow probability. In contrast with the previous section, we do take the call duration into account. Consequently we will examine the impact of both burst level and call level effects. First an expression for the overflow probability will be derived. Then the choice of the probability distributions of both call durations and sojourn times in the on and off state will be discussed in detail. A numerical example with which the use of large deviations asymptotics is justified, is followed by several numerical results concerning the impact of the call and burst level distributions. Finally we will discuss the implications of our analysis on MBAC.

3.1 Transient probability taking into account the call holding times

In Section 2 the number of flows present in the system was held constant, and consequently the study focused on the burst level fluctuations. In this subsection we examine the interaction between the burst and call level. We concentrate on the procedures for the case of momentary load measurements, but extensions to the other measurement procedures of Section 2.4 are straightforward.

Suppose a new flow is accepted. Once having accepted this flow, the system has to deal with all traffic generated by the flows already present, increased by the traffic of the newly arrived

flow. When detecting congestion the most rigorous measure the system can take is rejecting all flows afterwards, since the service of present flows can not be interrupted. For this reason, the relevant probability to consider is the probability of the aggregate input rate reaching the link rate between the epoch ($t = 0$) of acceptance of the new flow and the epoch that all flows that were present at time $t = 0$ have left the system, not accepting any new request at some time $t > 0$, given the specific value of the load offered at time 0. This probability is a lower bound for the overflow probability in case the system takes another measure when congestion is detected.
We make two simplifying assumptions. (i) We assume that all flows are of the on/off type. (ii) The probability under consideration in this section requires us to work with $n - 1$ flows that have their *residual* holding times ahead of them and 1 flow that has just been accepted. Instead we will assume that all n flows were present at time 0. This makes it easier to obtain an expression for the probability under consideration, as all n flows have the same residual holding time distribution. If the number n of calls in the system is large, the holding time of the new call is very unlikely to have any influence on the overflow probability. Hence, for the case when n is large, both expressions will nearly coincide.
As said above, we assume that n flows are present at time 0, of which $n\alpha$ are in the on state. The link rate C is denoted by $n\beta$ (corresponding with the notation introduced in Section 2). Without loss of generality, we assume the flows' peak rates to equal 1.

Our reasoning will be of a heuristic nature. As justified in [27], we will extensively use the so-called Laplace's principle [12], stating that the decay rate of an integral equals the decay rate of the maximum of the integrand.
We denote by $A_n(t)$ the number of flows (out of n) in the on-state at time t, assuming that they do not leave the system. Notice that this definition slightly differs from the one in Section 2, where $A_n(t)$ is a d-dimensional vector. Also, we define $\tilde{A}_n(t)$ to be the number of flows (out of the n that were present at time 0) in the on state at time t, taking into account that the flows leave the system after their holding time. Then the probability of interest can be written as

$$\tilde{q}_n(\alpha, \beta) := \int_t \mathrm{P}(\tilde{A}_n(t) = n\beta | A_n(0) = n\alpha). \tag{11}$$

First we will condition on the number of flows (out of the n flows that were present at time 0) that are still present at time t, $B_n(t)$. The residual holding time of a flow has distribution function $G(\cdot)$, which trivially yields

$$\mathrm{P}(B_n(t) = n\delta) = \binom{n}{n\delta} G(t)^{n-n\delta}(1 - G(t))^{n\delta}.$$

Using the independence of $B_n(t)$ and $A_n(0)$ we obtain

$$\mathrm{P}(\tilde{A}_n(t) = n\beta | A_n(0) = n\alpha, B_n(t) = n\delta) = \frac{\mathrm{P}(\tilde{A}_n(t) = n\beta, A_n(0) = n\alpha)}{\mathrm{P}(A_n(0) = n\alpha)}.$$

By conditioning on $A_{n\delta}(0)$, with $n\delta$ the number of flows present at time t this is equal to

$$\sum_{l=0}^{\min\{n\delta, n\alpha\}} \mathrm{P}(A_{n\delta}(t) = n\beta | A_{n\delta}(0) = l)\mathrm{P}(A_{n\delta}(0) = l | A_n(0) = n\alpha).$$

Since the number of flows (out of the $n\delta$ flows present at time t) in the on state at time 0 conditional on the number of flows present at time 0, follows a hypergeometric distribution, we have

$$\mathrm{P}(A_{n\delta}(0) = l | A_n(0) = n\alpha) = \frac{\binom{n\alpha}{l}\binom{n-n\alpha}{n\delta-l}}{\binom{n}{n\delta}}.$$

Conclude that $\tilde{q}_n(\alpha, \beta)$ roughly equals

$$\int_t \mathrm{P}(\tilde{A}_n(t) = n\beta | A_n(0) = n\alpha, B_n(t) = n\delta)\mathrm{P}(B_n(t) = n\delta) =$$

$$\int_t \sum_{n\delta, n\alpha'} \mathrm{P}(A_{n\delta}(t) = n\beta | A_{n\delta}(0) = n\alpha') \binom{n\alpha}{n\alpha'} \binom{n-n\alpha}{n\delta-n\alpha'} G(t)^{n-n\delta}(1-G(t))^{n\delta}. \tag{12}$$

The decay rate of this expression is equal to the decay rate of the maximum of the summands ('Laplace'). Now observe that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathrm{P}(A_{n\delta}(t) = n\beta | A_{n\delta}(0) = n\alpha') =$$

$$\delta \lim_{m\to\infty} \frac{1}{m} \log \mathrm{P}\left(A_m(t) = m\frac{\beta}{\delta} | A_m(0) = m\frac{\alpha'}{\delta}\right). \tag{13}$$

The decay rate of the right hand side of (13) is just the one identified in Section 2.2. Also, using Strirling's formula, the decay rate of the product of binomial coefficients in (12) equals

$$\alpha' \log\left(\frac{\alpha}{\alpha'}\right) + (\alpha - \alpha') \log\left(\frac{\alpha}{\alpha - \alpha'}\right)$$

$$+(\delta - \alpha') \log\left(\frac{1-\alpha}{\delta - \alpha'}\right) + (1 - \alpha - \delta + \alpha') \log\left(\frac{1-\alpha}{1-\alpha-\delta+\alpha'}\right). \tag{14}$$

Finally, it is evident that

$$\lim_{n\to\infty} \frac{1}{n} \log\left[G(t)^{n-n\delta}(1-G(t))^{n\delta}\right] = (1-\delta)\log G(t) + \delta \log(1-G(t)). \tag{15}$$

The sum of (13), (14), and (15) has to be maximized over all $t > 0$, $\delta \in [\beta, 1]$, and

$$\alpha' \in [\max\{0, \alpha + \delta - 1\}, \min\{\delta, \alpha\}].$$

The optimizing arguments – t^*, δ^*, and α'^* – have the following 'large deviations interpretation'. In the first place, t^* is the most likely epoch of overflow, in that, given that the input rate exceeds the link rate it happens with overwhelming probability at time t^* (n large). Then the most likely number of flows that is still present is $n\delta^*$. Among these $n\delta^*$ flows, there are $n\alpha'^*$ that were on at time 0.

Remark 3.1. In some cases, one will only have a measurement of the number $n\alpha$ of calls in the on state, without knowing the total number n of calls present. The above procedure can easily be extended to this case. Define $A(t)$ and $\tilde{A}(t)$ analogously to $A_n(t)$ and $\tilde{A}_n(t)$, but now for the case that the number of calls is unknown. It can be verified easily that, measuring the number $n\alpha$ of calls in the on state, the number of calls present in the system has a negative

binomial distribution with success probability π_1 and required number of successes $n\alpha$. Then the overflow probability reads

$$\int_t \sum_j \mathrm{P}(\tilde{A}_j(t) = n\beta | A_j(0) = n\alpha) \binom{j-1}{n\alpha - 1} \pi_0^{j-n\alpha} \pi_1^j.$$

Using techniques similar to the ones above, its decay rate equals

$$\max_{\alpha^o, t} \left(\lim_{n\to\infty} \frac{1}{n} \log \mathrm{P}(\tilde{A}_{n\alpha^o}(t) = n\beta | A_{n\alpha^o}(0) = n\alpha) + \right.$$
$$\left. \alpha \log\left(\frac{\pi_1 \alpha^o}{\alpha}\right) + (\alpha^o - \alpha) \log\left(\frac{\pi_0 \alpha^o}{\alpha^o - \alpha}\right) \right).$$

The first decay rate can be calculated as above; the optimization is over all α^o larger than α.

3.2 Choice of the probability distributions

In this section we will explain what kind of distributions we choose for the on and off times and for the call durations.

On and off times. In this paper, we consider for the on and off times two essentially different types of distributions: an exponential distribution, and a heavy-tailed distribution. We choose these types as they are in some sense opposite.
The exponential distribution is characterized by a 'fast decaying' tail; consequently the superposition of flows with exponential on and off times exhibits a short-range dependent behavior. Flows with exponential on and off times, allowing for rather explicit analysis [3, 20, 23], are therefore not appropriate to capture self-similar phenomena, which are known to apply for a broad range of applications, see e.g. [21] for Ethernet traffic.
In order to incorporate long-range dependence, distributions are considered with a non-exponential, *heavy* tail. As introduced in Boxma [4], such a tail can be modelled as the distribution being a regularly varying function. It is known that a superposition of calls of which the on time has a regularly varying tail of index ν between 1 and 2 shows long-range dependent behavior. In that case, the first moment of the on time exists, whereas the second moment is infinite. In our analysis of MBAC algorithms, we consider both the possibility of the on times and the off times having a heavy-tailed distribution.
As pointed out in Section 2.5, for our analysis techniques it is convenient to have on and off times of which an explicit LT is known. Unfortunately, this is not the case for the most 'straightforward' example of a heavy-tailed distribution, the Pareto distribution. Therefore, we use a particular heavy-tailed distribution of index $\nu = 1\frac{1}{2}$, identified in [5], with the following explicit expression for the LT

$$\beta(s) = 1 - \delta \frac{s}{(\sqrt{c} + \sqrt{s})^2}.$$

This distribution is characterized by two parameters; because we choose the parameter δ that relates to the probability mass in 0 equal to 1, the other parameter c is the reciprocal of the mean.

Holding times. As explained in Section 3.1, not only distributions on burst level (the on and off times of the flows) but also distributions on call level (the holding times) can influence the

	exp on & exp off		exp on & ht off	
t	LD	exact	LD	exact
0.05	$4.57 \cdot 10^{-23}$	$4.72 \cdot 10^{-24}$	$4.26 \cdot 10^{-22}$	$4.39 \cdot 10^{-23}$
0.10	$7.20 \cdot 10^{-17}$	$7.36 \cdot 10^{-18}$	$5.88 \cdot 10^{-17}$	$6.01 \cdot 10^{-18}$
0.20	$6.23 \cdot 10^{-12}$	$6.18 \cdot 10^{-13}$	$6.08 \cdot 10^{-13}$	$6.10 \cdot 10^{-14}$
0.50	$5.06 \cdot 10^{-08}$	$4.59 \cdot 10^{-09}$	$1.66 \cdot 10^{-09}$	$1.58 \cdot 10^{-10}$
1.00	$6.49 \cdot 10^{-07}$	$5.54 \cdot 10^{-08}$	$4.20 \cdot 10^{-08}$	$3.82 \cdot 10^{-09}$
2.00	$1.06 \cdot 10^{-06}$	$8.90 \cdot 10^{-08}$	$2.25 \cdot 10^{-07}$	$1.98 \cdot 10^{-08}$
5.00	$1.08 \cdot 10^{-06}$	$9.06 \cdot 10^{-08}$	$5.95 \cdot 10^{-07}$	$5.09 \cdot 10^{-08}$
10.0	$1.08 \cdot 10^{-06}$	$9.06 \cdot 10^{-08}$	$8.06 \cdot 10^{-07}$	$6.83 \cdot 10^{-08}$

Table 1: Probabilities of extreme fluctuations of the aggregate traffic rate, asymptotics and exact values for exp on times.

performance of the MBAC algorithm. Analogously to the choice of the distributions on burst level, we took both the exponential and a heavy-tailed distribution for the call holding times. In this case we only need the distribution of the residual holding times (instead of the LT) and therefore we can simply use the Pareto distribution with index ν between 1 and 2. Let X be a Pareto distributed random variable, $X \in [k, \infty]$, then its density f_X is known to be

$$f_X(x) = \frac{\nu k^\nu}{x^{\nu+1}}.$$

Notice that there is for given ν one degree of freedom left; consequently, one can fit the mean.

3.3 Performance of the large deviations asymptotics

In this section we will determine the probability of the traffic rate reaching the value $n\beta$ at time t (in seconds), $q_n(t, \alpha, \beta)$ (see (1)), in a numerical example for both exponentially and heavy-tailed distributed on and off times. Recall that we assume that calls do not leave the system. We will rely on the large deviations approximation of $q_n(t, \alpha, \beta)$, given by

$$\mathrm{P}(A_n(t) = n\beta | A_n(0) = n\alpha) \approx \exp(nI), \tag{16}$$

where decay rate I is the right hand side of (2). Since the exact value of $q_n(t, \alpha, \beta)$, formula (5), requires the factorial function to be applied to large numbers, it is difficult to determine this value numerically. To justify the use of approximation (16), we present here a comparison with the exact value (5) in a typical example for on/off flows.

We consider in our example speech, since it can be easily modeled as an on/off flow, with the on times corresponding to talk-spurts and off times corresponding to silences. We choose the mean on time equal to 1 s and the mean off time equal to 1.5 s in accordance with Schwartz [26, p. 26]; talk-spurts are usually in the order of 0.4 to 1.2 s and silences in the order of 0.6 to 1.8 s. Furthermore we assume 100 of these flows to be present at time 0 and 40 of them to be in their on state at time 0, i.e. $\alpha = 0.4$. Since the link can accommodate 66 flows that transmit on peak rate, overflow corresponds to $\beta = 0.66$.

In Table 1 and 2 the results are given for both the LD (large deviations) approximation and the exact value of $q_n(t, \alpha, \beta)$. The first column represents the case with exponential on and off times, the second column exponential on times and heavy-tailed off times, the third column

	ht on & exp off		ht on & ht off	
t	LD	exact	LD	exact
0.05	$4.99 \cdot 10^{-26}$	$5.16 \cdot 10^{-27}$	$1.96 \cdot 10^{-25}$	$2.02 \cdot 10^{-26}$
0.10	$1.96 \cdot 10^{-20}$	$2.02 \cdot 10^{-21}$	$1.15 \cdot 10^{-20}$	$1.18 \cdot 10^{-21}$
0.20	$8.81 \cdot 10^{-16}$	$8.97 \cdot 10^{-17}$	$1.09 \cdot 10^{-16}$	$1.12 \cdot 10^{-17}$
0.50	$2.35 \cdot 10^{-11}$	$2.31 \cdot 10^{-12}$	$1.10 \cdot 10^{-12}$	$1.10 \cdot 10^{-13}$
1.00	$2.81 \cdot 10^{-09}$	$2.65 \cdot 10^{-10}$	$1.50 \cdot 10^{-10}$	$1.46 \cdot 10^{-11}$
2.00	$4.74 \cdot 10^{-08}$	$4.30 \cdot 10^{-09}$	$4.59 \cdot 10^{-09}$	$4.32 \cdot 10^{-10}$
5.00	$3.02 \cdot 10^{-07}$	$2.63 \cdot 10^{-08}$	$7.39 \cdot 10^{-08}$	$6.65 \cdot 10^{-09}$
10.0	$5.70 \cdot 10^{-07}$	$4.88 \cdot 10^{-08}$	$2.43 \cdot 10^{-07}$	$2.13 \cdot 10^{-08}$

Table 2: Probabilities of extreme fluctuations of the aggregate traffic rate, asymptotics and exact values for ht on times.

heavy-tailed on times and exponential off times and the last column heavy-tailed on and off times. The LD approximation and the exact value differ for any value of t by one order in this example, which means that we can use the approximation instead of the exact expression. Since the transient probability $q_n(t, \alpha, \beta)$ obtained with (2) is for any t larger than the exact probability in this example, the overflow probability will be overestimated. Consequently it is very likely that any decision rule based on this probability will be conservative.
According to Sanov [27, p. 37], the decay rate of probability $q_n(t, \alpha, \beta)$ will converge to

$$\sum_{i=1}^{d} \beta_i \log\left(\frac{\beta_i}{\pi_i}\right).$$

In this example, the decay rate of $q_n(t, \alpha, \beta)$ will converge to -0.1374, which means that $q_n(t, \alpha, \beta)$ will converge to $1.08 \cdot 10^{-06}$. We see that this limit is reached fastest for the first situation where the on and off times are exponentially distributed. We will return to related phenomena in the next sections.

3.4 Impact of the distributions of on and off times

In this section, we investigate the influence of the shape of the distributions of bursts and silences. For different values of t (in seconds), we calculate the large deviations approximation of loss at time t taking into account that calls can leave the system. That is, we calculate the sum of (13), (14), and (15) maximized over δ and α' for different values of t. We take the same flows as in Section 3.3, meaning that we assume the peak rate to be 64 kbit/s, the link capacity 155 Mbit/s and the call durations to be exponential with mean 100 s. There are 5600 flows, of which 2240 are on at time 0.
In Table 3 and 4 the LD approximation and the decay rate of the sum of (13), (14), (15) maximized over δ and α', are given for the same on and off time distributions as in the previous section. We have also determined the most likely epoch t^* of the traffic rate reaching the link rate for those on and off time distributions. Our main conclusions are:

- The distributions of bursts and silences *do* have a significant impact. Strikingly, heavy-tailed bursts and silences are 'better' for MBAC than exponential ones, in that they yield a lower overflow probability. Recall the MBAC algorithm of Gibbens, Kelly and Key [15],

t	exp on & exp off LD $t^* = 0.8260$	decay rate	exp on & ht off LD $t^* = 0.9869$	decay rate
0.05	$3.76 \cdot 10^{-34}$	-0.0137	$2.04 \cdot 10^{-31}$	-0.0126
0.10	$1.20 \cdot 10^{-19}$	-0.0078	$8.19 \cdot 10^{-20}$	-0.0078
0.20	$4.04 \cdot 10^{-12}$	-0.0047	$1.65 \cdot 10^{-13}$	-0.0053
0.50	$4.36 \cdot 10^{-08}$	-0.0030	$9.04 \cdot 10^{-10}$	-0.0037
1.00	$1.15 \cdot 10^{-07}$	-0.0029	$4.74 \cdot 10^{-09}$	-0.0034
2.00	$5.68 \cdot 10^{-09}$	-0.0034	$6.83 \cdot 10^{-10}$	-0.0037
5.00	$1.87 \cdot 10^{-14}$	-0.0056	$5.21 \cdot 10^{-15}$	-0.0059
10.0	$2.62 \cdot 10^{-26}$	-0.0105	$8.94 \cdot 10^{-27}$	-0.0107
t^*	$1.34 \cdot 10^{-07}$	-0.0028	$4.74 \cdot 10^{-09}$	-0.0034

Table 3: Impact of distribution of bursts and silences; both the LD approximation of the probability and decay rate are given for exp on times. t^* denotes the most likely epoch of overflow.

that rejects a new call if the current measured load exceeds a threshold $s(n)$. The results of Table 3 and 4 can then be interpreted as follows. For flows with heavy-tailed on or off times, a higher value of $s(n)$ can be used. We return to this subject in Section 3.6. Apparently, the current state of the system is a 'better predictor' in case of heavy-tailed distributions. One might say that in that case a higher degree of positive correlation exists, cf. the long range dependent behavior of a superposition of a large number of flows with heavy-tailed on times.

- Related to the first conclusion, the following can be noticed regarding the 'most likely epoch of overflow'. Maximizing the sum of (13), (14), and (15) over δ, α' and t we have computed the most likely time period t^* at which the link rate will be exceeded by the input rate. The results in Table 3 and 4 show that for exponential bursts and silences the aggregate rate will sooner reach the link rate than for heavy-tailed bursts and silences.

3.5 Impact of the distribution of the call duration

We will now examine the role of the shape of the call holding time distribution. In order to do that we repeated the experiments of Table 3 and 4, but this time for Pareto distributed call durations (with mean 100 s). To make the call duration distributions more significant than the other heavy-tailed distributions involved (of burst and silences), we choose parameter $\nu = 1\frac{1}{4}$. Based on the following considerations, we expect the call duration distribution to play a role.

- In Section 3.1 we identified an expression for the decay rate of $\tilde{q}_n(\alpha, \beta)$. It in fact shows that overflow occurs as a combination of two effects. In the first place there is a *call effect*: flows may stay longer in the system than on average. In the second place there is a *burst effect*: the flows present send longer at a certain rate than usual. Loosely speaking, if the call duration is Pareto rather than exponential, there is a larger probability of the

t	ht on & exp off LD $t^* = 1.2629$	decay rate	ht on & ht off LD $t^* = 1.3618$	decay rate
0.05	$1.27 \cdot 10^{-43}$	-0.0176	$1.37 \cdot 10^{-41}$	-0.0168
0.10	$2.32 \cdot 10^{-27}$	-0.0110	$6.09 \cdot 10^{-28}$	-0.0112
0.20	$4.94 \cdot 10^{-18}$	-0.0071	$1.08 \cdot 10^{-19}$	-0.0078
0.50	$4.15 \cdot 10^{-12}$	-0.0047	$5.50 \cdot 10^{-14}$	-0.0055
1.00	$1.63 \cdot 10^{-10}$	-0.0040	$3.18 \cdot 10^{-12}$	-0.0047
2.00	$7.45 \cdot 10^{-11}$	-0.0042	$2.34 \cdot 10^{-12}$	-0.0048
5.00	$1.19 \cdot 10^{-15}$	-0.0061	$5.28 \cdot 10^{-17}$	-0.0067
10.0	$2.46 \cdot 10^{-27}$	-0.0109	$1.01 \cdot 10^{-28}$	-0.0115
t^*	$2.08 \cdot 10^{-10}$	-0.0040	$4.94 \cdot 10^{-12}$	-0.0046

Table 4: Impact of distribution of bursts and silences; both the LD approximation of the probability and decay rate are given for ht on times. t^* denotes the most likely epoch of overflow.

call having a large holding time, due to the heavy tail. Therefore, we would think that under the Pareto distribution the overflow probability $\tilde{q}_n(\alpha, \beta)$ is larger than under the exponential distribution.

- Consider the case of heavy-tailed on times. Suppose a flow is on at time 0; this means that the residual burst length has infinite mean; in case of exponential holding times (with a residual duration with finite mean) this means that very likely the flow stays in the on state the remainder of the call. In case of the Pareto holding times (with parameter $\nu = 1\frac{1}{4}$) the probability of the burst being ended seems larger (as the burst has $\nu = 1\frac{1}{2}$). Again this indicates a possible influence of the call holding time distribution on the overflow probability.

Surprisingly however, we got almost exactly the same table as in the exponential case. An explanation for that is given by the following Lemma (proven in the appendix).

Lemma 3.1 *Let $(X_k)_{k=1}^n$ be a sequence of positive i.i.d. random variables with $\mathrm{E}X < \infty$ and no probability mass in 0. Let $\hat{X}_k$ be the residual life of X_k, and $\hat{X}^{(\ell)}$ the ℓth order statistic of the residual lifetimes. Then*

$$\lim_{n\to\infty} \mathrm{P}\left(\hat{X}^{(\ell)} > \frac{x}{n}\right) = \sum_{r=0}^{\ell-1} \exp\left(-\frac{x}{\mathrm{E}X}\right) \frac{1}{r!} \left(\frac{x}{\mathrm{E}X}\right)^r .$$

In words, the lemma states that the distribution of the smallest among a large number of residual lifetimes $\hat{X}_k$ only depends on the distribution of the X_k through the mean $\mathrm{E}X$. In a similar fashion, a similar 'asymptotic insensitivity' can be shown for the expectation of the first order statistic; in particular: $n\hat{X}^{(1)} \to \mathrm{E}X$. These remarkable results have a number of implications:

- Although the residual call duration of any call present at time 0 has infinite mean under the Pareto distribution and finite mean under the exponential distribution, the first call

to leave has about the same finite mean in both situations (100/5600 s = 17.8 ms in the setting of Table 3 and 4). As the first calls to leave determine the 'call effect' identified above, based on the lemma one can expect that there is not such a difference as was suggested on the basis of the individual residual durations.

- *Irrespective of the distribution of the call duration*, the time until the first burst ends has mean of about 1/2240 s = 0.45 ms, whereas the first call clears after a time with mean 100/5600 s = 17.8 ms. So although the residual call duration of any call present at time 0 has a larger mean under the Pareto distribution than under the exponential distribution, the first call to leave has about the same mean, *which is much larger than the mean time until the first burst ends.* Consequently, there is a strong separation of time scales for the numbers used in our example: the process of calls leaving will only affect the overflow probability through the mean call duration. We believe that this property holds more generally: as long as there are orders of magnitude difference between the mean on and off times on the one hand, and the mean call duration on the other hand, the separation holds, regardless of the underlying distributions.

An other angle to look at the influence of the call duration is the following. As argued above, overflow occurs by a combination of two effects: the burst effect and the call effect. Now consider the optimizing arguments t^*, δ^*, and α'^*. Our numerical experiments show that for our specific parameters it holds that $G(t^*) \approx 1 - \delta^*$. In other words: the number of flows that have remained at the most likely epoch of overflow is almost equal to the mean number of flows that is present after t^* units of time. Consequently, in our example overflow is *not* due to long call durations (call effect), but rather to flows being in their on state more than could be expected on the basis of their equilibrium distribution (burst effect).

3.6 Conclusions regarding MBAC

As mentioned in the introduction, we are interested in the viability of an MBAC algorithm based on a simple threshold procedure. Mathematically, crucial in this procedure is the value of the threshold $s(n)$, i.e. the maximum number of calls in the on state at which a new call can be admitted. In other words, we are looking for the largest value of α in $\tilde{q}_n(\alpha, \beta)$, such that it does not exceed a certain value, in our example 10^{-6}. It is the sensitivity of α in the model's distributions that has our particular interest. In Table 5 the combinations of distributions of bursts, silences, and call lengths are listed. The parameters are the same as the ones underlying the results in Table 3 and 4. That is, 5600 calls are present at time 0, mean call duration equals 100 s, mean on time 1 s, mean off time 1.5 s, peak rate is equal to 64 kbit/s and the link capacity 155 Mbit/s.

Influence of the burst level distributions. In Table 5 we display the maximum value of α such that the overflow probability $\tilde{q}_n(\alpha, \beta)$ is small enough. It is seen that $s(n)$ for exponentially distributed bursts and silences is smaller than for heavy-tailed (see [5]). Hence, calls have to be rejected sooner for the situation where calls have exponentially distributed on and off times. At first sight the differences seem to be very small (with heavy-tailed on and off times roughly 40 calls more can be in their on time before arriving calls have to be rejected). However, it is evident that the difference can be made larger by taking bursts and silences with a distribution that decays faster than exponentially, and comparing it with heavy tailed bursts and silences with $\nu \in (1, 1\frac{1}{2})$.

burst level *call level*	exp on & exp off	ht on & ht off
exp	0.4075	0.4151
Pareto	0.4074	0.4150

Table 5: Maximum values for α such that $\tilde{q}_n(\alpha, \beta) < 10^{-6}$.

The relevance of this phenomenon can be further motivated as follows. Suppose the $s(n)$ is determined on the basis of heavy-tailed on/off traffic; according to the below table, this value is 2325. Using this value for exponential on/off traffic yields overflow probability $1.91 \cdot 10^{-05}$, about 20 times too large! In other words, the difference in $s(n)$ may not be that large, but the $s(n)$ curve is quite steep.
Consequently, it *is* important to take the distributions of the bursts and silences into account when determining a value for $s(n)$. Consequently, the approach of Gibbens, Kelly, and Key [15] (discussed in the introduction) will not lead to a safe MBAC algorithm for all realistic types of traffic. To develop a safe MBAC, one should base the $s(n)$ curve on the 'worst realistic' type of traffic. Therefore, an interesting question from a theoretical point of view is: what distribution of the on and off times (their means held fixed) maximizes the overflow probability, or, equivalently, minimizes threshold $s(n)$? Interestingly, Pareto bursts and silences are – from an MBAC point of view – not considered to be 'dangerous'.
Notice that this last conclusion only holds in the absence of buffers. If there are buffers the ordering may be different. Then again for exponential on/off sources the transient probabilities tend to the stationary probabilities faster than for heavy-tailed on/off sources. However, the stationary loss probability is now not insensitive to the distributions of bursts and silences, and typically larger in the case of heavy tails [4]. What effect is dominant is hard to predict, and requires knowledge of the transient of the corresponding queues.

Influence of the holding time distribution. We see that the value of $s(n)$ is not sensitive with respect to the call duration distribution. It should be noticed that consequently holding times with a heavy-tailed distribution do not negatively affect the MBAC: although there is a larger probability of calls remaining extremely long in the system, the only relevant parameter is the *mean* holding time.

4 Conclusion

In this paper we examined the impact of the distributions (on both the call and burst level) on MBAC algorithms with a threshold policy. To state it more precisely: we were interested in the influence of these distributions on the value of admission threshold of the momentary load, beyond which calls should be rejected. For a good MBAC algorithm, its performance should not critically depend on any traffic assumption.
To assess the robustness of the threshold based MBAC algorithm, we first derived a large deviations approximation for the transient overflow probability. As we assume the resource to be bufferless, this probability is defined as the probability that the input rate reaches the link rate at a certain time t, given the value of the input rate at time 0. This approximation is usually conservative, and numerically tractable. Its evaluation relies on the inversion of Laplace transforms.

We have shown that influence on the overflow probability of the burst level dynamics is significant. We compared flows with heavy-tailed bursts and silences with flows with exponentially distributed bursts and silences. An important insight is that calls of the former type are better in the sense that the overflow probability given a number of calls in the on state at time 0, is smaller. This phenomenon is also reflected in the value of the admission threshold: flows with exponential on and off times require a lower threshold. It is argued that an inaccurate choice of the threshold may lead to severe QoS violations. We conclude that if the choice of the threshold is based on transient overflow probabilities, the insensitivity to the burst level distributions does not apply, cf. [15].
Contrary to our expectation, the overflow probability and the threshold only depend on the distribution of the call holding time through its mean. We gave a mathematical justification of this phenomenon. It should be emphasized that this statement only holds under the condition that the mean call holding time is several orders of magnitude larger than the mean burst and silence ('separation of time scales'). This implies that heavy-tailed call durations do not harm the threshold based MBAC algorithm.
We stress that we only investigated the feasibility of MBAC mechanisms, and that we did not attempt to develop one ourselves. With this paper as a 'prestudy', we believe that the following steps could be taken. (i) Based on the above observations, we conclude that a threshold policy is a viable criterion. However, in order to find a threshold value for the measured momentary load that is safe for all (realistic) types of traffic, more research needs to be done. Then we have to identify which type of traffic is most 'aggressive', in that it demands the lowest admission threshold. One could for instance consider the class of on/off calls with given mean on and off time. (ii) We considered only homogeneous traffic feeding into the network. We believe that our main conclusions also hold in a heterogeneous environment, but this needs a more careful validation. (iii) An MBAC algorithm in general consists of both admission rules and measurement procedures; we hardly paid any attention to the latter. As stressed in the introduction, the measurements in this article were focused on capturing the current load, where we assumed that some basic traffic parameters (mean on and off times, mean call durations) were known from historic measurements. It should be noted that if one aims to estimate these parameters on-line, the procedure becomes more complicated: the uncertainty in the estimates has to be taken into account. Research in this direction is reported in [10, 16].

Appendix

Lemma 3.1 Let $(X_k)_{k=1}^n$ be a sequence of positive i.i.d. random variables with $\mathrm{E}X < \infty$ and no probability mass in 0. Let $\hat{X}_k$ be the residual life of X_k, and $\hat{X}^{(l)}$ the lth order statistic of the residual lifetimes. Then

$$\lim_{n\to\infty} \mathrm{P}\left(\hat{X}^{(l)} > \frac{x}{n}\right) = \sum_{r=0}^{l-1} \exp\left(-\frac{x}{\mathrm{E}X}\right) \frac{1}{r!} \left(\frac{x}{\mathrm{E}X}\right)^r.$$

Proof. Let $F(\cdot)$ be the distribution function of the X_k. First define the distribution function of the residual lifetimes (and its derivative):

$$Q(x) := \int_0^x \frac{1-F(u)}{\mathrm{E}X}\mathrm{d}u; \qquad Q'(x) = \frac{1-F(x)}{\mathrm{E}X}.$$

Using formula 4.2.21 in [2] and using the first order Taylor approximation, straightforward calculation yields

$$\begin{aligned}\lim_{n\to\infty} \mathrm{P}\left(\hat{X}^{(l)} > \frac{x}{n}\right) &= \lim_{n\to\infty} \sum_{r=0}^{l-1} \binom{n}{r} \left(1 - Q\left(\frac{x}{n}\right)\right)^{n-r} Q\left(\frac{x}{n}\right)^r \\ &= \sum_{r=0}^{l-1} \exp[-xQ'(0)] \frac{1}{r!} \left(xQ'(0)\right)^r \\ &= \sum_{r=0}^{l-1} \exp\left(-\frac{x}{\mathrm{E}X}\right) \frac{1}{r!} \left(\frac{x}{\mathrm{E}X}\right)^r,\end{aligned}$$

where the last equation is due to $F(0) = 0$.

□

Acknowledgments

During all stages of the preparation of this paper, Hans van den Berg (KPN Research) has given useful comments and suggestions for improvements.

References

[1] J. Abate and W. Whitt. The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems,* Vol. 10, 5 – 88, 1992.

[2] M. Abramowitz and I. Stegun. *Handbook of mathematical functions.* Dover, New York, USA, 1965.

[3] D. Anick, D. Mitra and M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *The Bell System Technical Journal,* 61: 1871 – 1894, 1982.

[4] O. Boxma. Fluid queues and regular variation. *Performance Evaluation,* 27 & 28: 699 – 712, 1996.

[5] O. Boxma and J. Cohen. The M/G/1 queue with heavy-tailed service time distribution. *IEEE Journal on Selected Areas in Communications,* 16: 749 – 763, 1998.

[6] A. Brandt and M. Brandt. On the distribution of the number of packets in the fluid flow approximation of packet arrival streams. *Queueing Systems*, 17: 275 – 315, 1994.

[7] F. Brichet and A. Simonian. Measurement-based CAC for video applications using SBR service. In: *Proceedings Performance and Management of Complex Communications Networks, IFIP 6.3 and 7.3*, Tsukuba, Japan, 1997.

[8] J. Cohen. Superimposed renewal processes and storage with gradual input. *Stochastic Processes and their Applications,* 2: 31 – 58, 1974.

[9] N. Duffield. Conditioned asymptotics for tail probabilities in large multiplexers. *Performance Evaluation,* 31: 281 – 300, 1998.

[10] N. Duffield. Asymptotic sampling properties of effective bandwidth estimation for admission control. In: *Proceedings Infocom*, 1532 – 1538, 1999.

[11] N. Duffield and W. Whitt. A source traffic model and its transient analysis for network control. To appear in: *Self-Similar Traffic and Performance Evaluation,* Wiley, New York, USA, 1999

[12] P. Dupuis and R. Ellis. *A weak convergence approach to the theory of large deviations.* Wiley, New York, USA, 1997.

[13] A. Elwalid, D. Mitra, and R. Wentworth. A new approach for allocating buffers and bandwidth to heterogeneous, regulated traffic in an ATM Node. *IEEE Journal on Selected Areas in Communications,* 13: 1115 – 1127, 1995.

[14] R. Gibbens and F. Kelly. Measurement-based connection admission control. In: *Proceedings 15th International Teletraffic Congress*, eds. V. Ramaswami and P. Wirth, 879 – 888, 1997.

[15] R. Gibbens, F. Kelly, and P. Key. A decision-theoretic approach to call admission control in ATM networks. *IEEE Journal on Selected Areas in Communications,* 13: 1101 – 1114, 1995.

[16] M. Grossglauser and D. Tse. A time-scale decomposition approach to measurement-based admission control. In: *Proceedings Infocom*, 1539 – 1547, 1999.

[17] L. Gün, T. Tedijanto, and P. Chimento. Dynamic connection admission mechanisms for the Networking BroadBand Services architecture. *Telecommunication Systems*, 7: 153 – 183, 1997.

[18] S. Jamin, P. Danzig, S. Shenker, and L. Zhang. A measurement-based admission control for integrated service packet networks. *IEEE/ACM Transactions on Networking*, 5: 56 – 70, 1997.

[19] G. Kesidis, J. Walrand, and C.-S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking*, 1: 424 – 428, 1993.

[20] L. Kosten. Stochastic theory of data-handling systems with groups of multiple sources. In H. Rudin and W. Bux (eds.), *Performance of Computer-Communication Systems*, 321 – 331, Elsevier Amsterdam, 1984.

[21] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Transactions on Networking*, 2: 1 – 15, 1994.

[22] M. Mandjes. Rare event analysis of the state frequencies of a large number of Markov chains. *Stochastic Models*, 15: 577 – 592, 1999.

[23] M. Mandjes and A. Ridder. Optimal trajectory to overflow in a queue fed by a large number of sources. *Queueing Systems*, 31: 137 – 170, 1999.

[24] D. Mitra, M. Reiman, and J. Wang. Robust dynamic admission control for unified cell and call QoS in statistical multiplexers. *IEEE Journal on Selected Areas in Communications,* 16: 692 – 707, 1998.

[25] J. Roberts, U. Mocci, and J. Virtamo (eds.) *Methods for the performance evaluation and design of broadband multiservice networks.* Springer, 1996.

[26] M. Schwartz. *Broadband Integrated Networks.* Prentice Hall, Upper Saddle River, USA, 1996.

[27] A. Shwartz and A. Weiss. *Large deviations for performance analysis, queues, communication, and computing.* Chapman and Hall, New York, USA, 1995.

[28] J. Wroclawski. Specification of the controlled load network element service. Internet Engineering Task Force RFC 2211, 1997.