Upper and Lower Bounds for the Multiplexing of Multiclass Markovian On/Off Sources

D. Artiges and P. Nain

INRIA RR-2734. Submitted to Performance Evaluation '96.

Abstract

In this paper, we consider a multiplexer with constant output rate and infinite buffer capacity fed by independent Markovian fluid on/off sources. We do not suppose that the model is symmetrical: there is an arbitrary number K of different traffic classes, and for each class k, an arbitrary number N_k of sources of this class. We derive lower and upper bounds for the stationary distribution of the backlog X of the form $B\exp(-\theta^\star x) \leq \mathrm{P}(X > x) \leq C(\theta)\exp(-\theta x)$, with $0 \leq \theta \leq \theta^\star$. When $K = 2$ or $K = 1$, we numerically compare our bounds to the exact distribution of X and to other previously known results. Through various examples, we discuss the behavior of $\mathrm{P}(X > x)$ and the tightness of the bounds.

Keywords: Exponential bound; Statistical multiplexing; Multiclass input; Markov fluid input; Effective bandwidth; Large deviation; Tail distribution.

1 Introduction

In a network node a large number of incoming data streams are multiplexed and share the common buffer space and bandwidth. New technologies, in particular ATM technology, make advantage of statistical multiplexing over more traditional multiplexing schemes (e.g. time division multiplexing, frequency division multiplexing) to allocate to each incoming stream a smaller output bandwidth than that that would be required if the input streams were all emitting at their maximum rate. The price to pay for this saving is the risk of overflow and congestion, which may harm network users in two different ways. Some of the traffic is time sensitive and suffers mainly from queueing delays building up when congestion occurs, while another part of the traffic may tolerate some delay but will not accept a single lost cell in the transmitted data. Estimating the delay and cell loss probability is thus an important part of network control. Two main problems arise: the first difficulty is to define mathematical models which render as closely as possible the principal characteristics of real traffic; the second difficulty is to analyze these models and derive accurate bounds or estimates, as in most cases getting exact results is out of reach either for computational or mathematical reasons.

In this paper, we address the situation when a multiplexer with constant service rate c and infinite buffer capacity is fed by $N = \sum_{k=1}^{K} N_k$ independent Markovian on/off Fluid Sources [1] (abbreviated as MFS in the following). More precisely, there are K classes of traffic and N_k MFS of class k. A source of class k emits data at a constant rate r_k when in state on and idles when in state off. The time spent by each source in the off (resp. on) state is exponentially distributed with mean

$1/\lambda_k$ (resp. $1/\mu_k$) for source k. Let $\hat{r} = \sum_k N_k r_k$ be the maximum instantaneous input rate, and $\bar{r} = \sum_k N_k \omega_k r_k$ be the mean input rate, where $\omega_k := \lambda_k/(\lambda_k + \mu_k)$ is the stationary probability that source k is in state on. We assume that $\bar{r} < c < \hat{r}$ so that the queue will immediately start to build up as soon as the total instantaneous rate exceeds c. Our objective is to find upper and lower bounds for the tail distribution of the stationary workload, $F(x) = \mathrm{P}(X > x)$, for all $x \geq 0$.

The model considered here, where the input data stream is represented as the superposition of a given number of MFS has been the subject of numerous studies in the past few years. In the *symmetrical* case ($K = 1$) Anick, Mitra and Sondhy in their pioneering work [1] showed that $F(x) = \mathbf{1}^T \sum_{i\,:\,\Re(z_i)<0} a_i\, \phi_i \exp(z_i\, x)$ for all $x \geq 0$, where (z_i, ϕ_i)'s are eigenvalue/eigenvector pairs, solutions to the eigenvalue problem $z\, \phi\, (\Lambda - c\, I) = \phi\, Q$, with Q the infinitesimal generator of the aggregate source and Λ the rate matrix. In the above spectral expansion, the coefficients a_i are obtained by solving a system of linear equations. This work was extended by Stern and Elwalid [31] to cover the case that the input stream is the superposition of reversible n-state Markov fluid sources. It is shown in [31] that $F(x)$ takes the same form as that obtained in the symmetrical case, where again some unknown numbers are solutions of a generalized eigenvalue problem. However, as pointed out by the authors, the exact computation of $F(x)$ may be infeasible for large systems. To overcome these difficulties, Stern and Elwalid investigated the case when the generalized eigenvalue problem decomposes into independent subsystems and devised an algorithm with a complexity of the order of $\prod_k N_k\, (\sum_k N_k)^3$ for computing $F(x)$ for all values of x (a brute force approach would require a complexity of the order of $\prod_k (N_k)^3$).

Approximations, bounds and asymptotics for $F(x)$ have also been recently proposed for handling cases when the computational burden for obtaining the exact value of $F(x)$ remains high, mainly when this computation has to be done in real-time, for instance, to perform control tasks like admission control of new sessions [11, 14, 18, 22].

In [29], Norros et al. used Beneš formula to derive an upper bound for $F(x)$ in the case when the input stream is the superposition of independent (non necessarily Markovian) on/off sources. However, computing the upper bound is not easy and various approximations of this upper bound have therefore been proposed by Norros et al. [29] and Bensaou et al. [4].

An important step in the analysis of statistical multiplexers was the discovery of the notion of *effective bandwidth* by Hui [23], later specialized to queueing models by Kelly [24], Gibbens and Hunt [17] and Guérin et al. [21]. Briefly, it has been noted that it is possible to associate an easily calculated quantity with each source of data, referred to as the effective bandwidth of that source, that captures the behavior of the tail of the response time at a multiplexer. In the case when the input traffic is the superposition of K types of N_k MFS of class k as described above, then this result takes the following form [14, 17]: $\lim_{x\to\infty} (1/x) \log \mathrm{P}(X > x) \leq -\theta$ if and only if $\sum_k N_k\, a_k(\theta) \leq c$, where $a_k(\theta)$, the effective bandwidth of a stream of type k is given by $a_k(\theta) = (\sqrt{g_k(\theta)^2 + 4\, \lambda_k\, r_k\, \theta} - g_k(\theta))/2\theta$ with $g_k(\theta) = \lambda_k + \mu_k - r_k\theta$. In matrix form, $a_k(\theta)$ appears to be the largest real eigenvalue of the matrix $Q_k + \theta\, \mathrm{diag}(0, r_k)$ from which it can be shown (see e.g. [25]) that $\theta\, a_k(\theta) = \lim_{t\to\infty} t^{-1} \log \mathbb{E}[\exp(\theta\, A_k(t))]$, where $A_k(t)$ is the amount of fluid generated by a source of type k in $[0, t)$. The latter result gives a nice interpretation of the effective bandwidth and shows that $\bar{r}_k \leq a_k(\theta) \leq \hat{r}_k$. Then, when the quality of service criterion is $\mathrm{P}(X > x) \leq \exp(-\theta\, x)$ for large x, call admission may be done by checking whether the effective bandwidth of the aggregate population, including the new stream seeking admittance, exceeds the service capacity. Note, however, that the effective bandwidth approach may be very conservative when used for small values of x [21, 28]. Other asymptotic results include the estimate for the loss probability for small buffers proposed by Hsu and Walrand [22], and various large deviation asymptotics in buffer statistics when the number of on-off sources gets large [13, 30, 33, 34].

In parallel to obtaining asymptotic bounds for $F(x)$, there have been efforts to get bounds for *any* value of x. In [5] Buffet and Duffield obtained an upper bound for $F(x)$ in the case of discrete and *homogeneous* MFS via a martingale approach. In [7] Chang developed a general theory based on Chernoff's inequality for computing exponential upper bounds, which allows him, in particular, to derive upper bounds for heterogeneous discrete time Markov on/off sources. In [28], Liu et al. obtained lower and upper bounds for $F(x)$ for a multiplexer fed by heterogeneous discrete time Markov on/off sources (see next section for further details). A non-exhaustive list of papers addressing the computation of exponential bounds for queueing models/multiplexers different than that considered in the present paper is [3, 9, 10, 12, 19, 26, 36].

In this paper we focus on Markov on/off fluid sources. The state of a class k source at time t is represented by a continuous time Markov process Y_t^k, with $Y_t^k = 0$ or 1 if the source is respectively silent or active at time t. The matrix Q_k, infinitesimal generator of Y_t^k is determined by the mean durations of active and silent periods of the source (resp. $1/\mu_k$ and $1/\lambda_k$):

$$Q_k = \begin{pmatrix} -\lambda_k & \lambda_k \\ \mu_k & -\mu_k \end{pmatrix}.$$

The Markov process $Y = (Y_t)_{t\in\mathbb{R}}$ describing the aggregate arrival stream is the cartesian product of the individual processes Y_t^k. The state space of Y is $\mathcal{S} = \{0,1\}^{\sum N_k}$, and its infinitesimal generator Q is the Kronecker sum (cf. [20] for definition, and [15]) of the matrices Q_k of each source:

$$Q = \bigoplus_{1\le k\le K} Q_k^{\oplus N_k}.$$

For all state s in $\mathcal{S}$, let $s_k^i = 0$ or 1 represent the state of the ith source of class k, let $s_k = \sum_i s_k^i$ be the number of active class k sources, and let $r(s) = \sum_k s_k r_k$ be the arrival rate in this state. From the assumption $\bar{r} < c < \hat{r}$, the stationary workload X exists and can be expressed as a function of the excess work arrived in the system in time intervals $(-t, 0]$:

$$X = \sup_{t\ge 0} \int_{-t}^{0} (r(Y_u) - c)du. \tag{1}$$

In the following, we use the above expression of the variable X to show that there exists a constant $\theta^\star > 0$, obtained as the unique solution in $(0,\infty)$ of the equation $\sum_k N_k\, a_k(\theta) = c$, such that, for $0 \le \theta \le \theta^\star$,

$$e^{-\theta^\star x} \prod_{k=1} \frac{\omega_k\, r_k}{a_k(\theta^\star)} \le F(x) \le C(\theta)\, e^{-\theta\, x}, \quad \forall x \ge 0. \tag{2}$$

In the above inequalities, the coefficient $\theta^\star$ is the exponential decay rate of the tail distribution of X, and provides asymptotical information for large x. But the asymptotical approximation $F(x) \approx \exp(-\theta^\star x)$ is generally far from accurate, and it is important to find the best possible coefficients $B = \prod_k \omega_k\, r_k / a_k(\theta^\star)$ and $C(\theta)$.

For heterogeneous MFS the coefficient $C(\theta)$ is obtained by solving an integer linear programming problem of complexity of the order of $\prod_k N_k$. We also provide an upper bound for $C(\theta)$ which can be computed by an algorithm of complexity of the order of $K\log(K)$, independent of the total numbers of sources.

For homogeneous MFS (we set $N_k = N$, $r_k = r$, $\lambda_k = \lambda$, $\mu_k = \mu$, $\omega_k = \omega$, $a_k(\cdot) = a(\cdot)$) the optimal exponential decay rate $\theta^\star$ and the coefficient $C(\theta^\star)$ are given by simple and explicit formulas. In this particular case, we retrieve with some improvements some previously known results.

The paper is organized as follows. The next section introduces some analytical tools and describes our approach to obtain the bounds in (2). This approach is developed in more details in Section 3, in which we give the algorithms to calculate the bounds. Some numerical results and an application to call admission control are presented in Section 4, before concluding.

2 Presentation of the approach

The bounds (2) are obtained by making use of some results of Liu, Nain, and Towsley in [28]. This section reviews rapidly the results from [28] which we need thereafter, and describes how they can be applied to our model.

2.1 Bounds for a process defined by recursion

In [28], Liu et al. extended the results of Kingman [26] and derived exponential lower and upper bounds on the tail distribution of a discrete time process Z_n satisfying the Lindley recursion

$$Z_{n+1} \quad = \quad [Z_n + U_n]^+ . \tag{3}$$

The statistical assumption on the sequence U_n is the following: there exists a Markov chain Y_n with finite state space $\mathcal{S}$ such that for all n, the pair (Y_{n+1}, U_n) conditioned on the process history $(Y_j, j \leq n; U_i, i \leq n-1)$ depends only on Y_n. However, for the use that we will make of the results in Section 3, we can simply assume that $U_n = f(Y_n)$ for some deterministic function f from $\mathcal{S}$ to $\mathbb{R}$, which is a special case of the previous condition. Let $P = [p_{st}]$ be the transition matrix of Y_n and $\pi = (\pi(s))_{s\in\mathcal{S}}$ be its stationary distribution, and define

$$\begin{aligned} \phi_s(\theta) &= \exp(\theta f(s)), \quad \text{for all } s \in \mathcal{S}, && (4)\\ \Phi(\theta) &= \operatorname{diag}(\phi_s(\theta); s \in \mathcal{S}), && (5)\\ H(\theta) &= \Phi(\theta)P. && (6) \end{aligned}$$

Let $\rho(\theta)$ be the spectral radius of the positive irreducible matrix $H(\theta)$, and let $z(\theta) = (z_s(\theta))_{s\in\mathcal{S}}$ be the corresponding positive left eigenvector of norm (sum of components) equal to 1. From [28, Lemma 2.2], there exists a stationary version Z of the process Z_n under the stability condition $\sum_s \pi(s)f(s) < 0$. Moreover, there is a unique positive and finite solution $\theta^\star$ to the equation

$$\rho(\theta) \quad = \quad 1, \tag{7}$$

and the distribution of Z satisfies the following inequalities: for all θ such that $0 \leq \theta \leq \theta^\star$,

$$B(\theta^\star)\exp(-\theta^\star x) \ \leq\ \mathrm{P}(Z > x) \ \leq\ C(\theta)\exp(-\theta x), \quad \forall x \geq 0, \tag{8}$$

where

$$
\begin{aligned}
B(\theta) &= \inf_{\substack{x>0 \\ s\in\mathcal{S}}} \frac{\sum_{t\in\mathcal{S}(x)} p_{ts}\, \pi(t)}{\sum_{t\in\mathcal{S}(x)} p_{ts}\, z_t(\theta)\, \exp(\theta(f(t)-x))}, && (9) \\
C(\theta) &= \sup_{\substack{x>0 \\ s\in\mathcal{S}}} \frac{\sum_{t\in\mathcal{S}(x)} p_{ts}\, \pi(t)}{\sum_{t\in\mathcal{S}(x)} p_{ts}\, z_t(\theta)\, \exp(\theta(f(t)-x))}, && (10) \\
\mathcal{S}(x) &= \{t \in \mathcal{S} \;/\; f(t) \geq x\}. && (11)
\end{aligned}
$$

We summarize in the next subsection the approach by discretization by which the above bounds apply to the present situation, before going into more details in Section 3.

2.2 Discretization approach

The main difficulty in our analysis comes from the fact that the bounds in (8) are defined for a discrete time process, characterized by a resursive equation, whereas the backlog process in the model defined in the introduction is a continuous time process that can be characterized by a differential equation.

We work this problem out in the following way: we split the time line into intervals of constant length δ, with $\delta > 0$ a discretization parameter meant to tend to 0, and we introduce a new process (X_n^δ, Y_n^δ), where the Markov chain Y_n^δ is the Markov process Y_t observed at time $n\delta$, for $n \in \mathbb{Z}$, and where X_n^δ is defined by a recursive equation similar to (3). The formulas (9) and (10) provide bounds on the tail distribution $\mathrm{P}(X^\delta > x)$ of the stationary version of X_n^δ. We first show that the variable X^δ tends almost surely to X as δ goes to 0. Then, we show that the bounds on $\mathrm{P}(X^\delta > x)$ also tend for all $x \geq 0$ to finite positive values as δ goes to 0. Finally, the desired bounds in (2) are obtained by taking the limit in (8) when $Z = X^\delta$.

Before proceeding with this method, we would like to make two comments. The first comment is to justify the use of the discretization approach for a continuous time model. Considering that we have to study a discrete time model at some point in order to make use of the bounds in (8), one might question the study of a continuous time model and the limiting scheme needed in our approach: why not just analyse a discrete time queueing model and directly apply the bounds from [28]? The fact is that we have calculated the bounds in (8) for a queueing model with discrete time Markovian on/off sources, but we have found that their expressions are complicated and difficult to compute. On the other hand, when taking the limit when the discretization parameter tends to 0, we observe that the expressions greatly simplify, and, as we will see below, the formulas which we obtain are easily implementable.

For many applications, and in particular for ATM networks, it would seem more appropriate to analyse discrete time models, as an ATM node inherently operates in a discrete time manner. However, the time unit in such a model is equal to the transmisison time of one cell, and is extremely small compared to the typical tolerable queueing delay: with a 155Mb/s outgoing link, this transmission time is about 3 microseconds, compared to tolerable queueing delays of the order of the millisecond or above. Thus, when trying to bound queueing delays or buffer overflow probabilities as we do here, the discrete time characteristic of an ATM multiplexer can be neglected and a continuous time model with fluid arrival makes perfect sense.

The second comment is to mention the existence of an alternative to the discretization approach. Instead of approximating X_t by a variable X_n^δ and letting δ go to 0, we can define a process (Z_n, Y_n), where Y_n is a Markov chain deduced from Y_t by uniformization (i.e. $Y_n = Y_{t_n}$ with (t_n) a Poisson process which includes all transition instants of Y_t) and where Z_n is equal to the process X_t observed at the uniformization times: $Z_n = X_{t_n}$. It can be shown that X_t and Z_n have the same stationary distribution. This uniformization approach provides a nice way to make use of the results from [28] and yields the same bounds as the discretization approach, but we do not develop it here because the proof relies on the characterization of $\mathrm{P}(X > x)$ through a differential equation and initial conditions (see [14]), which is still an unresolved conjecture. Details can be found in [2].

3 Derivation of the bounds

3.1 Discretization of the model and convergence

The discretized model is defined from the initial model by making the following changes. The Markov process Y_t is observed at times $n\delta$, for all $n \in \mathbb{Z}$, and the arrival rate of the input process is assumed to be constant in the time interval $[n\delta, (n+1)\delta)$ and equal to $r(Y_{n\delta})$. The amount of data entering the queue during this interval is thus $\delta r(Y_{n\delta})$. The service rate of the multiplexer is kept constant and equal to c as above, so that the amount of data transmitted in each interval of length δ is equal to $\delta\, c$ if the queue is not empty. If X_n^δ denotes the backlog of the system at time $n\delta$, we have

$$X_{n+1}^\delta = \left[X_n^\delta + \delta(r(Y_{n\delta}) - c)\right]^+ . \tag{12}$$

In this recursive equation, $(Y_{n\delta})_{n\in\mathbb{Z}}$ is a Markov chain on the state space $\mathcal{S}$ with transition matrix $P^\delta = e^{\delta Q}$. Note that $Y_{n\delta}$ has the same stationary distribution $(\pi(s))_{s\in\mathcal{S}}$ as Y_t. Let X^δ be the stationary regime of X_n^δ. We will present the bounds on $\mathrm{P}(X^\delta > x)$ in the next subsection, but we first show that the initial continuous time model can be considered as the limit of the discretized model when the parameter δ goes to 0. This convergence is expressed in the following lemma.

Lemma 3.1. *The variable X^δ tends almost surely to X as δ tends to 0.*

Proof We give some indications without going into details. Define $Y_t^\delta = Y_{n\delta}$ for all t such that $n\delta \leq t < (n+1)\delta$. Then,

$$X^\delta = \sup_{t\geq 0} \int_{-t}^{0} (r(Y_u^\delta) - c)du. \tag{13}$$

For all t, $|r(Y_t^\delta) - r(Y_t)|$ is no more than $\hat{r}$ and is zero on every interval $[n\delta, (n+1)\delta)$ where Y_t is constant, thus,

$$\int_{-t}^{0} (r(Y_u^\delta) - c)du \leq \int_{-t}^{0} (r(Y_u) - c)du + \delta\hat{r}(N(t) + 1), \tag{14}$$

with $N(t)$ the number of transitions of Y_t in $(-t, 0]$. From the ergodicity of Y_t, and from the fact that $N(t)/t$ tends almost surely to a finite positive number, it can be seen that for each sample path, there is a number $M > 0$ and a $\delta_0 > 0$ such that for all $t \geq M$ and $\delta < \delta_0$, both terms in (14) are negative. Then, from (1) and (13),

$$
\begin{aligned}
|X - X^\delta| &= \left| \sup_{0\le t\le M} \int_{-t}^{0} (r(Y_u) - c)du - \sup_{0\le t\le M} \int_{-t}^{0} (r(Y_u^\delta) - c)du \right| \\
&\le \sup_{0\le t\le M} \left| \int_{-t}^{0} (r(Y_u) - c)du - \int_{-t}^{0} (r(Y_u^\delta) - c)du \right| \\
&\le \delta \hat{r}(N(M) + 1).
\end{aligned}
$$

Letting $\delta \to 0$ in the above inequality yields the result. ∎

This convergence result will allow us to derive bounds on the probability $\mathrm{P}(X > x)$ by calculating bounds on $\mathrm{P}(X^\delta > x)$ with the inequalities (8) and by taking the limit as δ tends to 0.

3.2 Bounds in the discretized model

We calculate in this subsection the coefficients $\theta_\delta^\star$, $B_\delta(\theta)$, and $C_\delta(\theta)$ defining the bounds in (8) when $Z = X^\delta$. The exponential decay rate $\theta_\delta^\star$ is characterized by equation (7), we now try to give an explicit form to this equation. The transition matrix P^δ of the Markov chain $Y_{n\delta}$ is the Kronecker product, with exponents N_k, of the transition matrices $P_k^\delta = \exp(\delta Q_k)$ of individual discretized sources. We have

$$
P_k^\delta = \begin{pmatrix} 1 - p_k^\delta & p_k^\delta \\ q_k^\delta & 1 - q_k^\delta \end{pmatrix},
$$

and, for small δ,

$$
\begin{aligned}
p_k^\delta &= \delta\lambda_k + o(\delta), && (15) \\
q_k^\delta &= \delta\mu_k + o(\delta). && (16)
\end{aligned}
$$

We define the two-dimensional matrices

$$
\begin{aligned}
\Psi_k^\delta(\theta) &= \mathrm{diag}(1, \exp(\delta\theta r_k)), && (17) \\
J_k^\delta(\theta) &= \Psi_k^\delta(\theta) P_k^\delta. && (18)
\end{aligned}
$$

Let $\tau_k(\theta)$ be the spectral radius of the positive matrix $J_k^\delta(\theta)$, which is equal to its larger eigenvalue, and let $z_k^\delta(\theta)$ be the corresponding positive left eigenvector with components summing to 1. We further define the matrices of dimension $\#\mathcal{S}$

$$
\begin{aligned}
\Psi^\delta(\theta) &= \mathrm{diag}(\exp(\delta\theta r(s)), s \in \mathcal{S}), \\
J^\delta(\theta) &= \Psi^\delta(\theta) P^\delta,
\end{aligned}
$$

and we let $\tau^\delta(\theta)$ be the spectral radius of the positive matrix $J^\delta(\theta)$ and $z^\delta(\theta)$ be the corresponding positive left eigenvector with norm 1. From (7), the decay rate $\theta_\delta^\star$ is defined as the unique positive solution of the following equation:

$$
\tau^\delta(\theta_\delta^\star) = \exp(\delta\theta_\delta^\star c). \qquad (19)
$$

The matrices $\Psi^\delta(\theta)$ and $J^\delta(\theta)$ are the Kronecker products with powers N_k of the matrices $\Psi_k^\delta(\theta)$ and $J_k^\delta(\theta)$ respectively. As these matrices are positive, we have the following property on their spectral radii and eigenvectors (cf. [20]):

$$\tau^\delta(\theta) = \prod_{1\le k\le K} \tau_k^\delta(\theta)^{N_k}, \tag{20}$$

$$z^\delta(\theta) = \bigotimes_{1\le k\le K} z_k^\delta(\theta)^{\otimes N_k}. \tag{21}$$

Equations (20) and (21) provide an easy way to obtain $\tau^\delta(\theta)$ and $z^\delta(\theta)$: because each matrix $J_k^\delta(\theta)$ has only dimension 2, the spectral radius $\tau_k^\delta(\theta)$ and eigenvector $z_k^\delta(\theta)$ can be calculated explicitly with no difficulty. This property will be useful in Subsections 3.3 and 3.4 where we will have to take the limits of $\tau^\delta(\theta)$ and $z^\delta(\theta)$ as δ tends to 0.

We now express the coefficients in the bounds for the discretized model, by using formulas (9) and (10) of Subsection 2.1. We denote by $z_s^\delta(\theta_\delta^\star)$ the components of vector $z^\delta(\theta_\delta^\star)$, and by $p^\delta(t,s)$ the (t,s)-entry of the transition matrix P^δ. Then, the coefficients $B_\delta(\theta)$ and $C_\delta(\theta)$ in the inequalities (8) for $Z = X^\delta$ are

$$B_\delta(\theta) = \inf_{\substack{x>0\\ s\in\mathcal{S}}} F(s,x,\theta)\,, \qquad C_\delta(\theta) = \sup_{\substack{x>0\\ s\in\mathcal{S}}} F(s,x,\theta), \tag{22}$$

where

$$F(s,x,\theta) = \frac{\sum_{t\in\mathcal{S}(x)} p^\delta(t,s)\,\pi_t}{\sum_{t\in\mathcal{S}(x)} p^\delta(t,s)\,z_t^\delta(\theta)\exp(\theta(\delta(r(t)-c)-x))}, \tag{23}$$

and $\mathcal{S}(x) = \{t\in\mathcal{S} \;/\; x\le \delta(r(t)-c)\}$. The above formulas are derived directly from (9) and (10) with the remark that $f(s) = \delta(r(s)-c)$. We now wish to simplify the above expressions and define some other coefficients $B'_\delta(\theta)$ and $C'_\delta(\theta)$ such that $B'_\delta(\theta) \le B_\delta(\theta)$ and $C_\delta(\theta) \le C'_\delta(\theta)$, and which will be easier to compute. In the denominator of the fraction $F(s,x,\theta)$, note that

$$1 \le \exp(\theta(\delta(r(t)-c)-x)) \le \exp(\theta\delta\hat{r}),$$

and thus

$$\exp(-\theta\delta\hat{r})\frac{\sum_{t\in\mathcal{S}(x)} p^\delta(t,s)\,\pi_t}{\sum_{t\in\mathcal{S}(x)} p^\delta(t,s)\,z_t^\delta(\theta)} \le F(s,x,\theta) \le \frac{\sum_{t\in\mathcal{S}(x)} p^\delta(t,s)\,\pi_t}{\sum_{t\in\mathcal{S}(x)} p^\delta(t,s)\,z_t^\delta(\theta)}. \tag{24}$$

From the definition of the set $\mathcal{S}(x)$, the numerator and denominator of each fraction in (24) are constant functions of x on every interval $(a,b]$, where a and b are two consecutive values of $\delta(r(h)-c)$ with $h\in\mathcal{S}$. Thus, the extrema of the fraction over $x>0$ are equal to the extrema over the values of x of the form $\delta(r(h)-c)$, where h belongs to the set $\mathcal{T} = \{s\in\mathcal{S} \;/\; r(s)>c\}$. Define

$$B'_\delta(\theta) = \min_{\substack{h\in\mathcal{T}\\ s\in\mathcal{S}}} \frac{\sum_{r(t)\ge r(h)} p^\delta(t,s)\,\pi_t}{\sum_{r(t)\ge r(h)} p^\delta(t,s)\,z_t^\delta(\theta)}\exp(-\theta\delta\hat{r}), \tag{25}$$

$$C'_\delta(\theta) = \max_{\substack{h\in\mathcal{T}\\ s\in\mathcal{S}}} \frac{\sum_{r(t)\ge r(h)} p^\delta(t,s)\,\pi_t}{\sum_{r(t)\ge r(h)} p^\delta(t,s)\,z_t^\delta(\theta)}. \tag{26}$$

Then, we have from (24)

$$B'_\delta(\theta) \;\le\; \inf_{\substack{x>0\\ s\in\mathcal{S}}} F(s,x,\theta), \qquad \text{and} \qquad C'_\delta(\theta) \;\ge\; \sup_{\substack{x>0\\ s\in\mathcal{S}}} F(s,x,\theta),$$

and thus $B'_\delta(\theta) \le B_\delta(\theta) \le C_\delta(\theta) \le C'_\delta(\theta)$.

It is easier to deal with $B'_\delta(\theta)$ and $C'_\delta(\theta)$ than with $B_\delta(\theta)$ and $C_\delta(\theta)$ because the extremum is taken over a finite set in (25) and (26); this property is used in Subsection 3.4 below. Note that the ratio between $B'_\delta(\theta)$ and $C'_\delta(\theta)$ or between $C'_\delta(\theta)$ and $C_\delta(\theta)$ is no more than $\exp(\theta\delta\hat{r})$. As this value tends to 1 for all θ when δ goes to 0, and as our objective is to take the limit of the bounds in δ, we see that there is no loss of tightness by considering the simplified expressions: if the limits of the coefficients $B_\delta(\theta)$ and $C_\delta(\theta)$ exist, then the limits of $B'_\delta(\theta)$ and $C'_\delta(\theta)$ also exist and are the same.

3.3 Exponential decay rate $\theta^\star$ and effective bandwidth

We now show that the decay rate $\theta^\star_\delta$ defined in equation (19) for the discretized model has a finite positive limit when δ goes to 0. The first step is to establish the following Taylor expansion of the spectral radius $\tau_k(\theta)$:

$$\tau_k^\delta(\theta) \;=\; 1 + \delta\theta a_k(\theta) \;+\; o(\delta), \tag{27}$$

where the function $a_k(\theta)$ is defined by

$$a_k(\theta) \;=\; \frac{1}{2\theta}\left(\sqrt{(\lambda_k+\mu_k+r_k\theta)^2 - 4\mu_k r_k\theta} - (\lambda_k+\mu_k-r_k\theta)\right). \tag{28}$$

The proof of (27) is as follows: the spectral radius $\tau_k(\theta)$ of the matrix $J_k^\delta(\theta)$ is its larger eigenvalue, which is calculated explicitly as the matrix has only dimension 2. Then, the expansion is obtained through some straightforward calculus, by using (15) and (16).

From (20) and (27), we can write

$$\tau^\delta(\theta) \;=\; 1 + \delta\theta a(\theta) \;+\; o(\delta), \tag{29}$$

with

$$a(\theta) \;=\; \sum_k N_k a_k(\theta). \tag{30}$$

It is easily shown that the function $a(\theta)$ is strictly increasing and varies continuously from $\bar{r} = a(0)$ to $\hat{r} = \lim_{\theta\to+\infty} a(\theta)$. From the inequality $\bar{r} < c < \hat{r}$, we deduce that the equation

$$a(\theta) \;=\; c \tag{31}$$

has a unique positive solution, which we call $\theta^\star$. From the properties of the function $a(\theta)$, this solution satisfies the following characterization: for all $\theta > 0$,

$$\theta \;<\; \theta^\star \quad\Longleftrightarrow\quad G(\theta) \;<\; \theta c, \tag{32}$$

$$\theta \;>\; \theta^\star \quad\Longleftrightarrow\quad G(\theta) \;>\; \theta c. \tag{33}$$

From a similar characterization of the number $\theta_\delta^\star$ due to the log-convexity of $\tau^\delta(\theta)$ (see [28, Lemma 2.1]), and from the expansion (29), we obtain the limit

$$\lim_{\delta\to 0} \theta_\delta^\star \quad = \quad \theta^\star. \tag{34}$$

Thus, we have shown that the decay rate $\theta_\delta^\star$ in the discretized model tends to a number $\theta^\star$, which is the unique positive solution of equation (31).

As mentioned in the introduction, the function $a(\theta)$ is the effective bandwidth of the arrival stream, and is equal to the sum of the effective bandwidths $a_k(\theta)$ of all sources. We see from (31) that the effective bandwidth $a(\theta)$ of a stream in a multiplexer represents the share of bandwidth which is consumed by this stream when the workload distribution in the multiplexer has decay rate θ. The general formula for this function is (see e.g. [25]):

$$a(\theta) \quad = \quad \limsup_{t\to+\infty} \frac{1}{\theta\, t} \log \mathbb{E}[\exp(\theta A(t))], \tag{35}$$

with $A(t)$ the amount of data generated by the stream in $[0,t)$. The effective bandwidth of a MFS lies between its mean and peak rates, but we now illustrate on an example the influence of the parameters λ_k and μ_k on the function $a_k(\theta)$. The example is with two classes of MFS: sources of class 1 have parameters $1/\lambda_1 = 8$s, $1/\mu_1 = 2$s, and $r_1 = 1$Mb/s; and sources of class 2 have parameters $1/\lambda_2 = 0.8$s, $1/\mu_2 = 0.2$s, and $r_2 = 1$Mb/s. Class 1 MFS is burstier than class 2 MFS in the sense that the active periods, when the source is emitting at peak rate, are on the average 10 times longer for class 1 than for class 2, but the sources of both classes have the same mean and peak rates. Figure 1 compares the effective bandwidth $a(\theta)$ as a function of θ for three aggregate streams, each composed of 32 sources of class 1 and 2 in different proportions. Although these streams have all the same peak rate 32Mb/s and same mean rate 6.4Mb/s, we observe that their effective bandwidths are very different. This example shows the important influence of the duration of a burst on the effective bandwidth function.

This concludes the study of the decay rate $\theta^\star$. The next subsection presents the calculation of the prefactor coefficients in the bounds.

3.4 Coefficients of the upper and lower bounds

We show that the coefficients $B_\delta(\theta_\delta^\star)$ and $C_\delta(\theta_\delta^\star)$ in the discretized model have positive limits B and C as the discretization parameter δ tends to 0. For sake of simplicity, we present only the derivation of the coefficients $B = B(\theta^\star)$ and $C = C(\theta^\star)$. An upper bound of the form $C(\theta)\exp(-\theta x)$ with a parameter $\theta < \theta^\star$ can also be obtained, where the coefficient $C(\theta)$ is worked out by simply replacing $\theta^\star$ and $\theta_\delta^\star$ by θ in the following.

We first introduce some notation. Let

$$u_{k,0} \quad = \quad \frac{1}{2(\lambda_k+\mu_k)} \left(\sqrt{(\lambda_k+\mu_k+r_k\theta^\star)^2 - 4\mu_k r_k \theta^\star} + (\lambda_k+\mu_k+r_k\theta^\star) \right), \tag{36}$$

$$u_{k,1} \quad = \quad \frac{1}{2(\lambda_k+\mu_k)} \left(\sqrt{(\lambda_k+\mu_k+r_k\theta^\star)^2 - 4\mu_k r_k \theta^\star} + (\lambda_k+\mu_k-r_k\theta^\star) \right). \tag{37}$$

It is easily seen that $0 < u_{k,1} < 1 < u_{k,0}$. The coefficients B and C are based on these numbers, which we can also express as a function of the effective bandwidth $a_k(\theta)$ of a class k source, of the

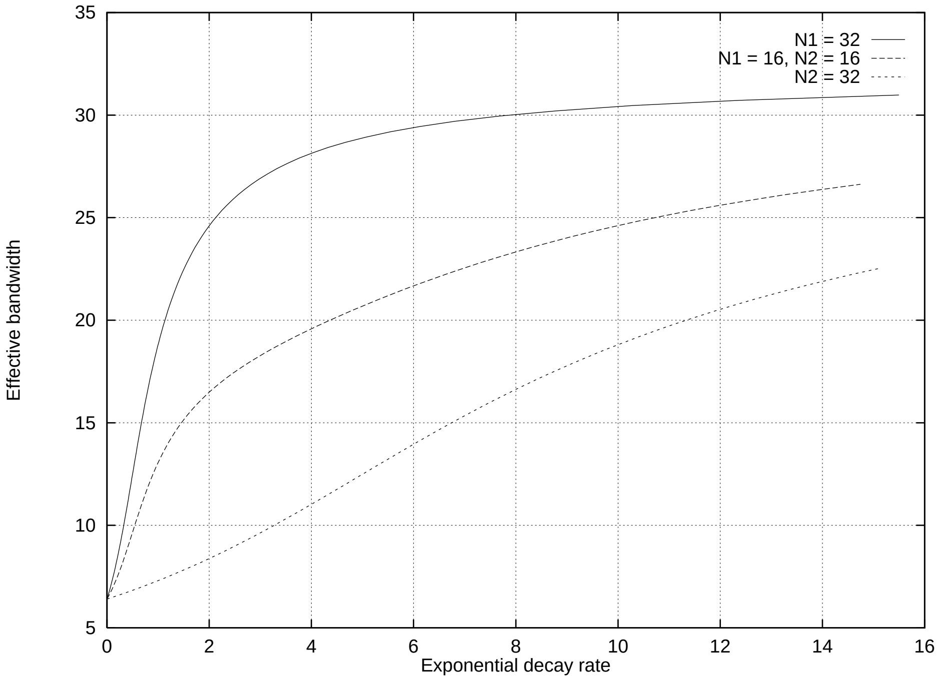

Figure 1: Influence of burst duration on the effective bandwidth.

mean active period duration $b_T = 1/\mu_k$, and of the peak and mean rates r_k and $\bar{r}_k = \omega_k r_k$. The formulas are easily obtained and are given without proof:

$$u_{k,0} = 1 + \theta^\star b_T \left(1 - \frac{\bar{r}_k}{r_k}\right) a_k(\theta^\star), \tag{38}$$

$$u_{k,1} = \frac{\bar{r}_k}{a_k(\theta^\star)}. \tag{39}$$

For all $t \in \mathcal{S}$, recall that t_k is the number of active class k sources in state t, and define

$$u(t) = \prod_{1 \le k \le K} u_{k,0}^{N_k - t_k} u_{k,1}^{t_k}. \tag{40}$$

We now calculate the limit of the components of vector $z^\delta(\theta^\star_\delta)$, which appear in the expression of B_δ and C_δ. For all t in $\mathcal{S}$, we have

$$\lim_{\delta \to 0} z_t^\delta(\theta^\star_\delta) = \pi_t / u(t). \tag{41}$$

The above formula is proved in the following way. If we denote by $z_{k,0}^\delta(\theta^\star_\delta)$ and $z_{k,1}^\delta(\theta^\star_\delta)$ the components of the two-dimensional vector $z_k^\delta(\theta^\star_\delta)$, then, we have from (21)

$$z_t^\delta(\theta^\star_\delta) = \prod_k z_{k,0}^\delta(\theta^\star_\delta)^{N_k - t_k} z_{k,1}^\delta(\theta^\star_\delta)^{t_k}. \tag{42}$$

Furthermore, as the different sources are independent, the stationary distribution π is such that

$$\pi_t = \prod_k (\bar{\omega}_k)^{N_k - t_k} (\omega_k)^{t_k}, \tag{43}$$

where we have defined $\bar{\omega}_k = 1 - \omega_k$. By comparing (40), (42), and (43), we see that the limit (41) is established if we prove that, for all k,

$$\lim_{\delta \to 0} z^{\delta}_{k,0}(\theta^{\star}_{\delta}) = \bar{\omega}_k / u_{k,0}, \tag{44}$$

$$\lim_{\delta \to 0} z^{\delta}_{k,1}(\theta^{\star}_{\delta}) = \omega_k / u_{k,1}. \tag{45}$$

We show the previous limit formulas by calculating the eigenvector $(z^{\delta}_{k,0}(\theta^{\star}_{\delta}), z^{\delta}_{k,1}(\theta^{\star}_{\delta}))$ of $J^{\delta}_k(\theta^{\star}_{\delta})$, through a linear system of dimension 2, as a function of the spectral radius $\tau^{\delta}_k(\theta^{\star}_{\delta})$ and of the components of the matrix $J^{\delta}_k(\theta^{\star}_{\delta})$. The equalities (44) and (45) follow after some elementary algebra by using the expansion formulas (15), (16), and (27).

We now calculate an expansion of the coefficients $p^{\delta}(t,s)$ of P^{δ}, the transition matrix of the Markov chain $(Y_{n\delta})_{n \in \mathbb{Z}}$. By using the independence of the sources, we can write explicitly the probability $p^{\delta}(t,s)$ in terms of the individual probabilities p^{δ}_k and q^{δ}_k: let n^k_{00} be the number of class k sources which are silent in both states t and s, n^k_{01} the number of class k sources which are silent in state t but active in state s, and n^k_{10} and n^k_{11} defined in a similar way. We have then, using again (15) and (16),

$$\begin{aligned} p^{\delta}(t,s) &= \prod_k (1 - p^{\delta}_k)^{n^k_{00}} (p^{\delta}_k)^{n^k_{01}} (1 - q^{\delta}_k)^{n^k_{11}} (q^{\delta}_k)^{n^k_{10}} &(46)\\ &= \prod_k (\delta \lambda_k)^{n^k_{01}} (\delta \mu_k)^{n^k_{10}} + o(\delta^{n^k_{01} + n^k_{10}}). &(47) \end{aligned}$$

For all pair (t,s) of $\mathcal{S}$, we denote by $|t-s|$ the total number of sources which are in a different state in s and t: $|t-s| = \sum_{k,i} |t^i_k - s^i_k|$, where t^i_k is 0 or 1 if the ith source of class k is respectively silent or active in state t, and similarly for s. By observing that $\sum_k n^k_{01} + n^k_{10} = |t-s|$, we deduce from (47) that there is a positive number $b(t,s)$, equal to $\prod_k \lambda_k^{n^k_{01}} \mu_k^{n^k_{10}}$, independent of δ such that

$$p^{\delta}(t,s) = b(t,s) \delta^{|t-s|} + o(\delta^{|t-s|}). \tag{48}$$

The limiting formulas (41) and (48) will now enable us to show that, despite their apparently complicated form, the coefficients $B'_{\delta}(\theta^{\star}_{\delta})$ and $C'_{\delta}(\theta^{\star}_{\delta})$ given by (25) and (26) respectively, and the coefficients $B_{\delta}(\theta^{\star}_{\delta})$ and $C_{\delta}(\theta^{\star}_{\delta})$, have calculable limits when δ tends to 0. We first show that $B'_{\delta}(\theta^{\star}_{\delta})$ tends to a positive number B when δ tends to 0. As noted above, this will show in the same time that the coefficient $B_{\delta}(\theta^{\star}_{\delta})$ also tends to B as δ goes to 0. Let $\mathcal{T} = \{s \in \mathcal{S} \ / \ r(s) > c\}$, and for all states h in $\mathcal{T}$ and s in $\mathcal{S}$, let

$$A_{\delta}(h,s) = \frac{\sum_{r(t) \geq r(h)} p^{\delta}(t,s)\, \pi_t}{\sum_{r(t) \geq r(h)} p^{\delta}(t,s)\, z^{\delta}_t(\theta^{\star}_{\delta})}. \tag{49}$$

We can calculate the limit of $A_{\delta}(h,s)$ when δ goes to 0 by considering two cases.

- 1st case: $r(s) \geq r(h)$. In this case, from (48), all terms of the sum in the numerator and denominator of $A_{\delta}(h,s)$ tend to 0 with δ, except the term corresponding to $t = s$, for which the limit of $p^{\delta}(t,s)$ is equal to 1. From the limiting formula (41), we see that $A_{\delta}(h,s)$ tends to the number $A(h,s)$ defined by

$$A(h,s) = u(s). \tag{50}$$

- 2nd case: $r(s) < r(h)$. Let i be the minimum of $|t-s|$ over all states t such that $r(t) \geq r(h)$, and let $\mathcal{S}(h,s) = \{t \in \mathcal{S} \ / \ r(t) \geq r(h) \text{ and } |t-s| = i\}$. From (41) and (48), the number $A_\delta(h,s)$ tends to a limit $A(h,s)$ as δ tends to 0, with

$$A(h,s) \quad = \quad \frac{\sum_{t\in\mathcal{S}(h,s)} b(t,s)\,\pi_t}{\sum_{t\in\mathcal{S}(h,s)} b(t,s)\,\pi_t/u(t)}. \tag{51}$$

From (25), we have $B'_\delta(\theta^\star_\delta) = \exp(-\theta^\star_\delta \delta \hat{r}) \times \min\{A_\delta(h,s) \ / \ h \in \mathcal{T}, s \in \mathcal{S}\}$. When δ goes to 0, $A_\delta(h,s)$ tends to $A(h,s)$ for all pair (h,s) with h in $\mathcal{T}$ and s in $\mathcal{S}$, thus, the minimum of $A_\delta(h,s)$ tends to the minimum of $A(h,s)$, because the minimization is done over a finite number of pairs (h,s). The term $\exp(-\theta^\star_\delta \delta \hat{r})$ has limit 1 and thus, $B'_\delta(\theta^\star_\delta)$ tends to the number B defined by

$$B \quad = \quad \inf\{A(h,s) \ / \ h \in \mathcal{T}, s \in \mathcal{S}\}. \tag{52}$$

The minimum of $A(h,s)$ is easily obtained: let $\hat{s}$ be the state of $\mathcal{S}$ in which all sources are active ($\hat{s}_k = N_k$ pour tout k). As $u_{k,1} < 1 < u_{k,0}$, we have for all state t the inequality $u(t) \ \geq \ u(\hat{s})$, and thus $A(h,s) \ \geq \ u(\hat{s})$ for all pair (h,s), which leads to $B \ \geq \ u(\hat{s})$. On the other hand, the state $\hat{s}$ is in $\mathcal{T}$, for $r(\hat{s}) = \hat{r} > c$, and thus $B \ \leq \ A(\hat{s},\hat{s}) = u(\hat{s})$. We have finally $B = u(\hat{s})$, which we can rewrite, from the definition of $u(s)$ in (40),

$$B \quad = \quad \prod_{1\leq k\leq K} u_{k,1}^{N_k}. \tag{53}$$

In a similar manner, one can show that the coefficient $C'_\delta(\theta^\star_\delta)$ defined in (26) tends to a positive number C as δ tends to 0, with

$$C \quad = \quad \sup\{A(h,s) \ / \ h \in \mathcal{T}, s \in \mathcal{S}\}.$$

Let m be a state in the set $\mathcal{T}$ such that $u(m)$ is maximum. This state exists because the set $\mathcal{T}$ is finite. Then, from the definition of $A(h,s)$, we have $A(h,s) \ \leq \ u(m)$ for all $h \in \mathcal{T}$ and all $s \in \mathcal{S}$, and thus $C \ \leq \ u(m)$. But we also have $C \ \geq \ A(m,m) = u(m)$, because m is in $\mathcal{T}$, which yields

$$C \quad = \quad u(m). \tag{54}$$

Thus, in order to derive the upper bound coefficient C, it is necessary to calculate $u(m)$, which is equivalent to find integer numbers m_k satisfying the following conditions:

$$0 \leq m_k \leq N_k, \tag{55}$$

$$\sum_{1\leq k\leq K} m_k r_k > c, \tag{56}$$

$$\prod_{1\leq k\leq K} u_{k,0}^{N_k - m_k} u_{k,1}^{m_k} \quad \text{is maximum.} \tag{57}$$

We can reformulate the problem of calculating C as a problem of integer programming of the knapsack type: let $m'_k = N_k - m_k$, then the above conditions are equivalent to

$$0 \leq m'_k \leq N_k, \tag{58}$$

$$\sum_{1\leq k\leq K} m'_k r_k < \hat{r} - c, \tag{59}$$

$$\sum_{1\leq k\leq K} m'_k \log(u_{k,0}/u_{k,1}) \quad \text{is maximum.} \tag{60}$$

Note that for all k, $\log(u_{k,0}/u_{k,1}) > 0$. The coefficient C is then given by the formula

$$\log(C) \quad = \quad \log(B) + \sum_{1 \le k \le K} m'_k \log(u_{k,0}/u_{k,1}). \tag{61}$$

In terms of complexity, the exact calculation of the coefficient C is a difficult problem: the simplest algorithm uses a dynamic programming approach where all possible combinations of numbers m'_k are considered, and has a complexity of $O(\prod_k N_k)$ (see Garfinkel and Nemhauser [16]). To overcome this difficulty, we consider the equivalent problem in real numbers: let D be the number defined by

$$\log(D) \quad = \quad \log(B) + \sum_{1 \le k \le K} x_k \log(u_{k,0}/u_{k,1}), \tag{62}$$

where the x_k are real numbers satisfying $0 \le x_k \le N_k$, $\sum_k x_k r_k \le \hat{r} - c$, and such that $\sum_k x_k \log(u_{k,0}/u_{k,1})$ is maximum.

Because the numbers m'_k have to satisfy stronger conditions than the x_k, we have clearly $C \le D$, thus $D \exp(-\theta^\star x)$ is also an upper bound of $\mathrm{P}(X > x)$, although possibly not as tight as $C \exp(\theta^\star x)$. The gain in considering the coefficient D instead of C is that the computational complexity is much lower: assume that the classes are sorted in decreasing order of $\log(u_{k,0}/u_{k,1})/r_k$ (the sorting requires $O(K \log K)$ computations), then the numbers x_k are given by the following straightforward algorithm, of complexity $O(K)$. Initial parameters are Volume $:= \hat{r} - c$, Value $:= \log(B)$, $x_i := 0$ for all i, and $k := 1$.

While (Volume > 0)
  do
    $x_k := \min(N_k, \text{Volume}/r_k)$
    Volume $:=$ Volume $- x_k r_k$
    Value $:=$ Value $+ x_k \log(u_{k,0}/u_{k,1})$
    $k := k + 1$
  done

The condition $0 < \hat{r} - c < \sum_k N_k r_k$ ensures the correct termination of the algorithm, and $\log(D)$ is given by the parameter Value at the exit of the while loop. The numerical comparisons in section 4 show that the difference between the bounds $\log(C)$ and $\log(D)$ is relatively small, which suggests that the improvement of tightness from one bound to the other is not worth the extra computational cost.

We obtain upper bound coefficients $C(\theta)$ and $D(\theta)$ with non optimal parameter $\theta < \theta^\star$ through the same algorithm as C and D, by simply replacing $\theta^\star$ with θ in the formulas for $u_{k,0}$ and $u_{k,1}$.

3.5 Summary of results

We now summarize our analytical results. The model considered is the statistical multiplexing of Markovian fluid on/off sources. We have defined a function $a(\theta)$, the effective bandwidth of the arrival process, and have shown that the exponential decay rate $\theta^\star$ is the unique positive solution of the equation $a(\theta) = c$. This characterization of the decay rate through the function $a(\theta)$ was presented before in [14], with some development.

We have then introduced three coefficients B ($B = B(\theta^\star)$), $C(\theta)$, and $D(\theta)$ which define exponential lower and upper bounds on the tail distribution $\mathrm{P}(X > x)$. The coefficient B has an explicit form (assuming $\theta^\star$ is known) and is easily calculated, $\log(C(\theta))$ is the value of an integer programming problem, and $\log(D(\theta))$ is obtained through a simple and fast algorithm. The bounds are the following: for all $x \geq 0$, and for all $0 \leq \theta \leq \theta^\star$,

$$B \exp(-\theta^\star x) \;\leq\; \mathrm{P}(X > x) \;\leq\; C(\theta) \exp(-\theta x) \;\leq\; D(\theta) \exp(-\theta x). \tag{63}$$

As mentioned in the introduction, these bounds generalize some previous results on this model in several ways. First, the inequalities (63) hold for any x, which is an important property when considering for example overflow probabilities for small buffers. This generalizes asymptotic results for large x. Secondly, the bounds hold for any number K of different traffic classes, which generalizes a number of other results for homogeneous sources, and they also hold for any number N_k of sources of any type, which generalizes some asymptotic results for a large number of sources. Furthermore, the bounds are easily calculable numerically. If $\theta^\star$ is known, the derivation of the lower bound coefficient B is immediate with formula (53). But note that the calculation of $\theta^\star$, by solving numerically the equation $a(\theta) = c$, is not required to obtain the upper bounds in (63): for a given $\theta > 0$, the condition $\theta \leq \theta^\star$ is equivalent to $a(\theta) \leq c$, and the coefficients $C(\theta)$ or $D(\theta)$ are derived directly from the effective bandwidth $a_k(\theta)$ of the sources and simple traffic parameters (peak rate, mean rate, burst duration). Thus, the calculation of an upper bound, say $D(\theta)\exp(-\theta x)$, does not necessitate more than some standard traffic information, plus the effective bandwidth $a(\theta)$, which we assume is implemented for CAC or tariffing purposes.

Finally, we remark that in some sense, the coefficients B and $D = D(\theta^\star)$ are the best possible. Consider the case where the service rate c is such that the only state s in $\mathcal{S}$ with input rate $r(s)$ larger than c is the state $\hat{s}$, where all sources are on. This condition is true if we have $\hat{r} - \min_k r_k < c < \hat{r}$. In this case, it is easily seen that $B = D$, and thus, the lower and upper bounds are equal and give the exact queue length distribution: $\mathrm{P}(X > x) = D \exp(-\theta^\star x)$. The bounds are then obviously the best possible. Although the study of this particular case is not necessarily of great practical importance, it suggests that the tightness of the coefficients may not be easily improved.

4 Application examples and experiments

We now illustrate our results through a number of numerical examples. In Subsection 4.1, we assume $K = 2$ and compare our bounds with the exact distribution of X. In Subsection 4.2, we assume $K = 1$ and observe that the bounds take a very simple form, and that the decay rate $\theta^\star$ has an explicit expression. Finally, in Subsection 4.3, we consider the problem of call admission control (CAC) in a network.

4.1 Numerical Comparisons when $K = 2$

For every system considered, the bounds were obtained by following the steps described in Subsection 3.3 and 3.4: we solved numerically the equation $a(\theta) = c$, calculated $\log(B)$, and derived $\log(C)$ through an integer programming algorithm. The exact backlog distribution was computed via the procedure described by Elwalid and Mitra in [14], by solving a linear system of differential equations. The difficult part in this procedure is to find all the eigenvalues and eigenvectors of a matrix of dimension $\prod_k (1 + N_k)$. This spectral analysis and the resolution of a linear system were implemented by using some functions of the library Meschach, a freeware package in C language for linear algebra (see reference manual [32]). In theory, the exact distribution can be obtained for any number of sources, but in practice, only very small systems can be studied because of the high computation time.

Figures 2, 3, and 4 show a few examples of the models that we have considered. Each has two different classes of traffic ($K = 2$) and a relatively small number of sources of each class ($N_k \leq 12$). The mean durations of off and on periods, respectively $1/\lambda_k$ and $1/\mu_k$, are given in seconds, and the rate r_k in Mb/s. In each figure, $\log_{10}(\mathrm{P}(X > x))$ is plotted as a function of the queueing delay in the multiplexer, which is equal to x/c, and compared to the lower and to both upper bounds. All three bounds are represented by straight lines with the same slope $\theta^\star$, with the value at the point $x = 0$ being equal to $\log_{10}(B)$, $\log_{10}(C)$, and $\log_{10}(D)$.

In Figure 2, the load $\bar{r}/c$ is 0.6, sources of class 1 represent voice channels, with parameters $1/\lambda_1 = 0.650$, $1/\mu_1 = 0.352$, $r_1 = 0.064$, and sources of class 2 model data streams with average on and off periods of 0.2 and 0.8 seconds respectively, and peak rate 320 Kb/s. In Figure 3, the parameters are the same but for the load which is taken equal to 0.40.

We first notice that in all cases that we have considered, $\log_{10}(\mathrm{P}(X > x))$ appears to be a nonincreasing and convex function of x. Thus, the gap between $\log_{10} \mathrm{P}(X > x)$ and the upper bound (that is a linear function of x) is the smallest for $x = 0$, whereas the gap between $\log_{10} \mathrm{P}(X > x)$ and the lower bound is the smallest for large values of x.

The second observation is that the lower bound can be very close to the exact value when x is not close to 0, and when the load of the system is very low, i.e. when the service rate c is of the same order as $\hat{r}$ (but still with $c < \hat{r} - \min_k r_k$). This phenomenon is illustrated in Figure 4 where we have taken $\bar{r}/c = 0.125$. However, for medium or high loads, the exact value is usually closer to the upper bound, even for larger x.

We further investigate decomposition properties of our bounds. Because of the difficulty to analyze exactly large systems, it would be interesting to know how to quickly and accurately extrapolate the behavior of $\mathrm{P}(X > x)$ with many sources from the study of smaller systems, where the exact distribution of X or some approximations of this distribution are more easily obtained.

Let d be a positive integer, which will act as a "scaling factor", let X_d be the stationary backlog in a system with dN_k class k sources and with a service rate equal to dc: the size of the system has been "multiplied" by d. How does $\mathrm{P}(X_d > dx)$ compare to $\mathrm{P}(X > x)$? We do not attempt here to bring a precise answer to this question, but a simple remark can be made about the scaling property of our bounds: we observe that the decay rate $\theta^\star$ does not depend on d (this comes from the equality $a_d(\theta) = da(\theta)$), and if B_d and D_d denote the coefficients in the scaled system, it is also easily shown that $\log(B_d) = d\log(B)$ and $\log(D_d) = d\log(D)$, which yields

$$d(\log(B) - \theta^\star x) \;\leq\; \log(\mathrm{P}(X_d > dx)) \;\leq\; d(\log(D) - \theta^\star x). \tag{64}$$

Thus, we can easily deduce the bounds for large systems from the bounds calculated for small systems.

The above formula also reveals a difficulty: our ability to estimate $\log(\mathrm{P}(X > x))$ is determined by the gap between the upper and the lower bound, which is equal to $\log(D) - \log(B)$. This difference grows linearly with the size of the system, so that we may expect that at least one of the bounds in (64) deviates from the exact curve when the number of sources is large. We continue the discussion on the behavior of the bounds in the next subsection, where we present numerical examples with a larger number of sources and with $K = 1$.

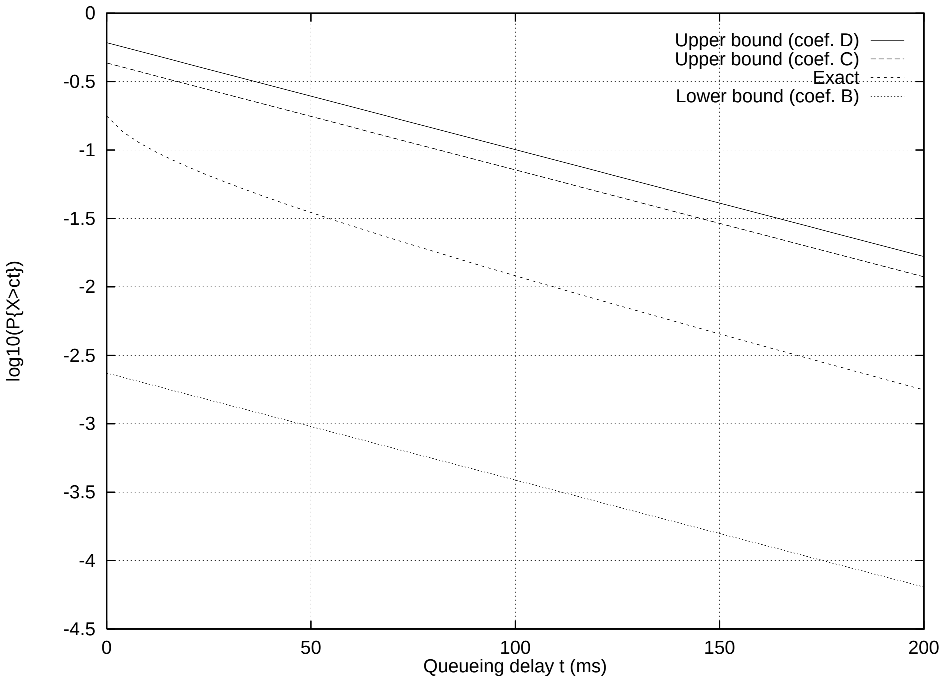

Figure 2: $(N_1, 1/\lambda_1, 1/\mu_1, r_1) = (12, 0.650, 0.352, 0.064)$, $(N_2, 1/\lambda_2, 1/\mu_2, r_2) = (6, 0.8, 0.2, 0.32)$, Load $= 0.6$.

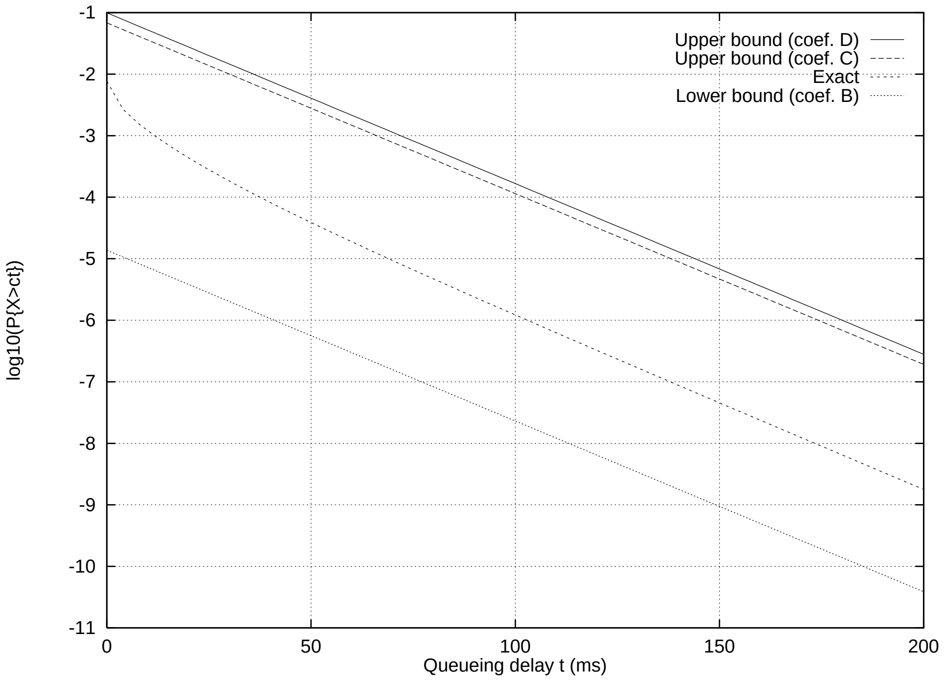

Figure 3: $(N_1, 1/\lambda_1, 1/\mu_1, r_1) = (12, 0.650, 0.352, 0.064)$, $(N_2, 1/\lambda_2, 1/\mu_2, r_2) = (6, 0.8, 0.2, 0.32)$, Load $= 0.4$.

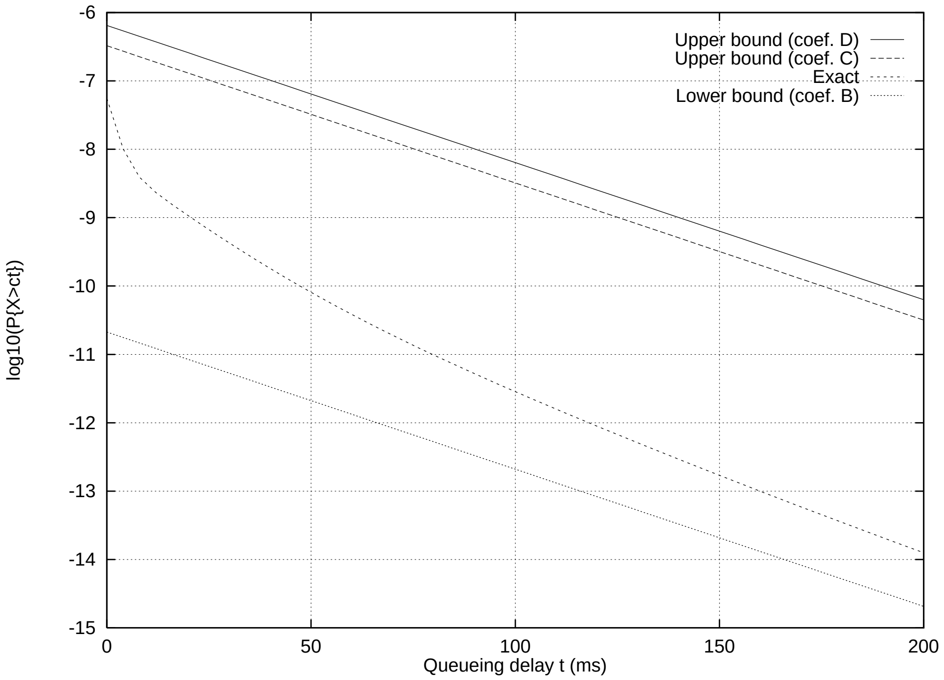

Figure 4: $(N_1, 1/\lambda_1, 1/\mu_1, r_1) = (12, 9.0, 1.0, 1.0)$,
$(N_2, 1/\lambda_2, 1/\mu_2, r_2) = (4, 9.0, 1.0, 2.0)$, Load$=0.125$.

4.2 Bounds in a symmetrical model ($K = 1$)

When $K = 1$, the simplicity of the formulas and the ability to confront them with other known results enable us to get a more precise insight on the behavior of the bounds. We consider a system with N sources of the same class, with parameters λ, μ, and r. We denote by c_1 the service rate per source, i.e. $c_1 = c/N$. In this case, the solution of the equation $a(\theta) = c$ is

$$\theta^\star = \frac{\mu}{r - c_1} - \frac{\lambda}{c_1}. \tag{65}$$

The above formula is not new and can be found for instance in [1] and [34]. If we report this value in (36) and (37) to calculate the numbers u_0 and u_1, we find:

$$\begin{aligned} u_0 &= \frac{\mu r}{(\lambda + \mu)(r - c_1)} \\ u_1 &= \frac{\lambda r}{(\lambda + \mu) c_1}. \end{aligned}$$

We now define $m = \lfloor c/r \rfloor + 1$, the minimum number of active sources such that the total input rate exceeds c. Note that $m > N c_1 / r$. Then, the coefficients of the bounds are

$$B = u_1^N, \tag{66}$$

$$C = u_0^{N-m} u_1^m, \tag{67}$$

$$D = (u_0^{1 - c_1/r} u_1^{c_1/r})^N. \tag{68}$$

We can write

$$D = \exp(-N I(c_1)), \tag{69}$$

with

$$I(c_1) \quad = \quad -(1 - c_1/r) \log(u_0) - (c_1/r) \log(u_1). \tag{70}$$

The above formula for the coefficient D was previously known as a large deviation approximation for the overflow probability in a bufferless model with large N (see for instance Weiss [34]), our work shows that this approximation is really an upper bound. This upper bound was also obtained by Buffet and Duffield in [5] for a discrete time model: we can derive D as the limit of their formula, if we let the discrete time model which they study tend to a continuous time model, as we have done in section 3.

We now comment on some numerical experiments. Figures 5 and 6 compare the logarithms of three different functions: the upper bound $C\exp(-\theta^\star x)$, the exact distribution $\mathrm{P}(X > x)$, and an approximation for small buffers due to Hsu and Walrand (see [22]). Each curve is plotted as a function of the queueing delay x/c in the multiplexer. The exact value of $\mathrm{P}(X > x)$ was computed by the method proposed by Anick, Mitra, and Sondhi in [1], which is not fundamentally different from the heterogeneous case, but leads to a simpler and quicker algorithm, which makes it possible to analyze larger systems. The approximation found in [22] is of the form $A(N)\exp(-NC_2\sqrt{x})$, where $A(N)$ is an estimate of the probability that the input rate exceeds c, and the coefficient C_2 is derived for small buffer asymptotics by Weiss [34].

We consider a system of 100 sources modelling voice channels ($1/\lambda = 0.650$, $1/\mu = 0.352$, $r = 0.064$). In Figure 5, the service rate c is such that the load of the system is 0.82, and in Figure 6, the load is 0.66. The buffer occupation X is represented by the corresponding queueing delay in milliseconds. On both figures, we observe that our upper bound is close to the real distribution for small x ($x = c\times$delay), but, as noted also in the heterogeneous model, the gap may increase by several orders of magnitude for larger x. The small buffer approximation is also very close for small x.

The lower bound with coefficient B is in both cases very inferior to the real probability ($\log(B)$ would be about -8 in Figure 5 for $x = 0$ and -18 in Figure 6), and was left out of the picture. This is not a surprise: as mentioned above, the difference between $\log(D)$ and $\log(B)$ grows linearly with the size N of the system and thus, at least one of the two bounds is expected to miss the exact value by a large margin when the number of sources increases.

When $K = 1$, we have seen that $\log(D)$ is the large deviation approximation of $\log(\mathrm{P}(X > 0))$ when N goes to $+\infty$, the difference between these two terms is thus $o(N)$, which is consistent with our observations: the upper bound $\log(D)$ is good when x is small, even for large N. This implies that $\log(\mathrm{P}(X > 0)) - \log(B)$ has to grow linearly with N. Thus, for large N, the upper bound is such that $\log(D)$ stays close to $\log(\mathrm{P}(X > 0))$, while the lower bound is such that $\log(B)$ deviates linearly in N from the exact value.

Let $\Phi(N)$ be the stationary probability that the input rate exceeds c, i.e. the probabiliy of having m or more active sources at one time, which is given by the formula

$$\Phi(N) \quad = \quad \sum_{k=m}^{N} \binom{N}{k} \left(\frac{\lambda}{\lambda+\mu}\right)^k \left(\frac{\mu}{\lambda+\mu}\right)^{N-k}. \tag{71}$$

We have

$$\Phi(N) \quad \leq \quad \mathrm{P}(X > 0), \tag{72}$$

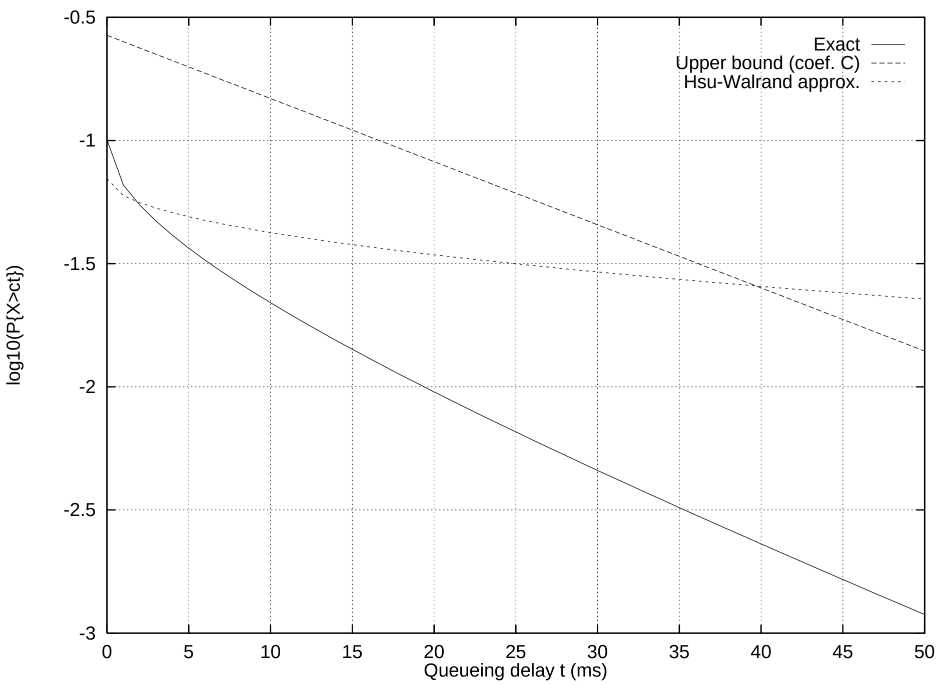

Figure 5: Load $= 0.82$, $N = 100$

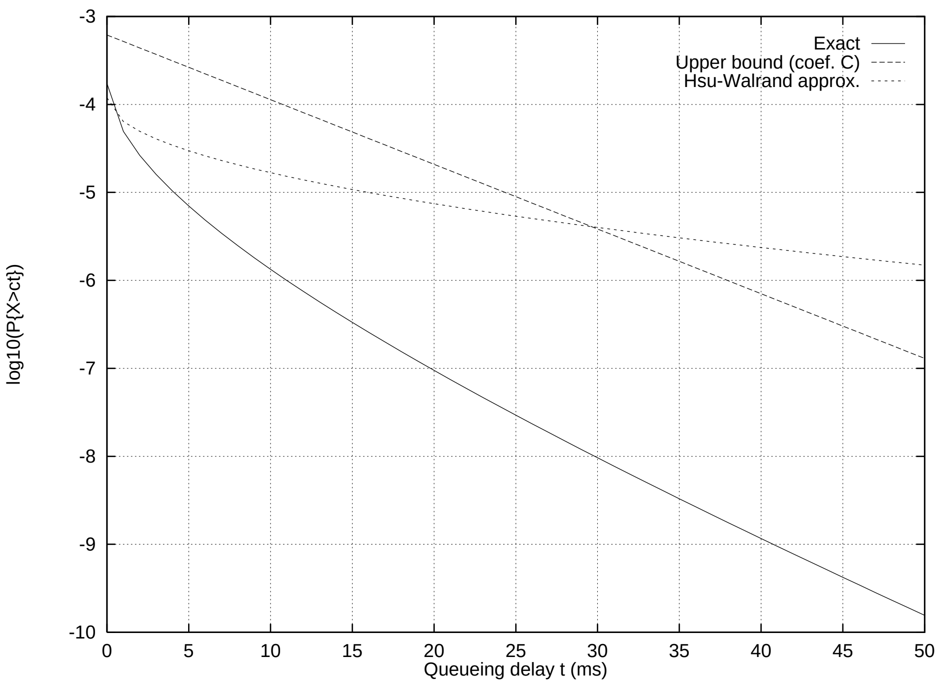

Figure 6: Load $= 0.66$, $N = 100$

because the queue length is positive whenever the input rate is larger than c, and thus $\Phi(N) \leq C$. The probability $\Phi(N)$ can be seen as an approximation of $\mathrm{P}(X > 0)$; because the formula (71) is complex, authors have searched approximations of $\Phi(N)$ itself. An approximation of $\log(\Phi(N))$, obtained by refinement of a large deviation formula, is presented in [22] and has the following form:

$$\log(A(N)) \quad = \quad -h(c_1) - \frac{1}{2}\log(N) - NI(c_1). \tag{73}$$

In Figures 7 and 8, we compare numerically $\log(\Phi(N))$, the approximation $\log(A(N))$, and the upper bound $\log(C)$, for up to 200 sources with two different sets of parameters. We can see on the curves that the approximation $\log(A(N))$ is very close to the exact value of $\log(\Phi(N))$, even for large N. Recall that $\log(D) = -NI(c_1)$. Thus, empirically, the difference between the upper bound $\log(D)$ and $\log(\Phi(N))$ is of the order of $\log(N)/2$. From the inequality (72), we finally deduce that the difference between $\log(D)$ and $\log(\mathrm{P}(X > 0))$ is also at most of the order of $\log(N)/2$.

This concludes our discussion on the behavior and accuracy of the bounds in the numerical examples: for a large number of sources, the lower bound with coefficient B is in general very inferior to the exact distribution of X, whereas the upper bound, with coefficient C or D, stays reasonably close.

4.3 Application to call admission control

The problem of call admission control has been one of the principal applications of the recent developments in the theory of effective bandwidths. Papers of interest dealing with this subject include [6], [14], [8], and [18], among many others.

We now briefly describe the simple mechanism based on the effective bandwidth formula. It relies on two assumptions. The first assumption is that the effective bandwidth $a_k(\theta)$ of each source is known. The second assumption is that the QoS criterion is uniform over all the connections and has the form $\mathrm{P}(X > b) \leq q$, for some predefined parameters b and q.

The exponential decay rate $\theta^\star$ of the tail distribution of X is larger than θ if and only if the effective bandwidth $a(\theta)$ of the arrival process is such that $a(\theta) \leq c$. The CAC mechanism uses the approximation $\mathrm{P}(X > x) \approx \exp(-\theta^\star x)$: the QoS criterion is then equivalent to the condition

$$a(\theta_0) \quad \leq \quad c, \tag{74}$$

where θ_0 is such that $\exp(-\theta_0 b) = q$. The decision to accept or reject a $(J+1)$st connection is taken by calculating the effective bandwidth $a(\theta_0) + a_{J+1}(\theta_0)$ of the input stream with the new connection. If this value stays below c, then the algorithm considers that the connection $J+1$ can be accepted without violating the condition $\mathrm{P}(X > b) \leq q$, otherwise, the connection attempt is rejected. We now illustrate this CAC mechanism through a numerical example, and show how our bounds can improve the performances.

The question is to find the maximum number N of fluid Markovian sources with parameters $1/\lambda = 0.650$s, $1/\mu = 0.352$s, and $r = 0.064$Mb/s, which can be accommodated in a multiplexer with $c = 3.2$Mb/s such that $\mathrm{P}(X > b) \leq q$? We can derive four different numbers:

- Exact number N_e: with the formulas proposed by [1], the exact distribution of X is known for all N. We obtain N_e by calculating the largest N such that $\mathrm{P}(X > b)$ does not exceed q.

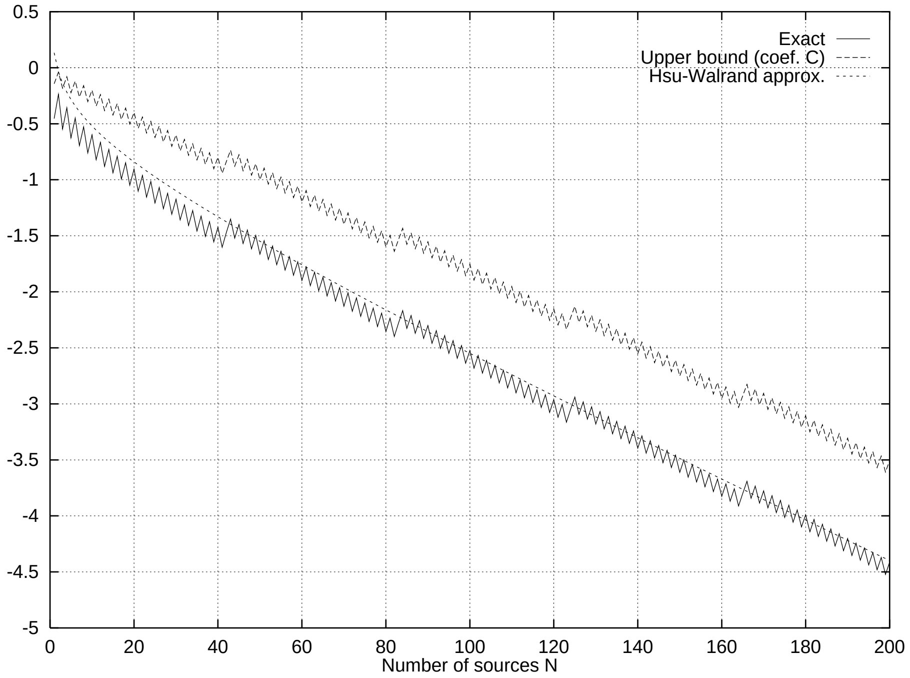

Figure 7: Plot of $\log_{10}(\Phi(N))$, with load=0.72 and $(1/\lambda, 1/\mu, r) = (0.650, 0.352, 0.064)$

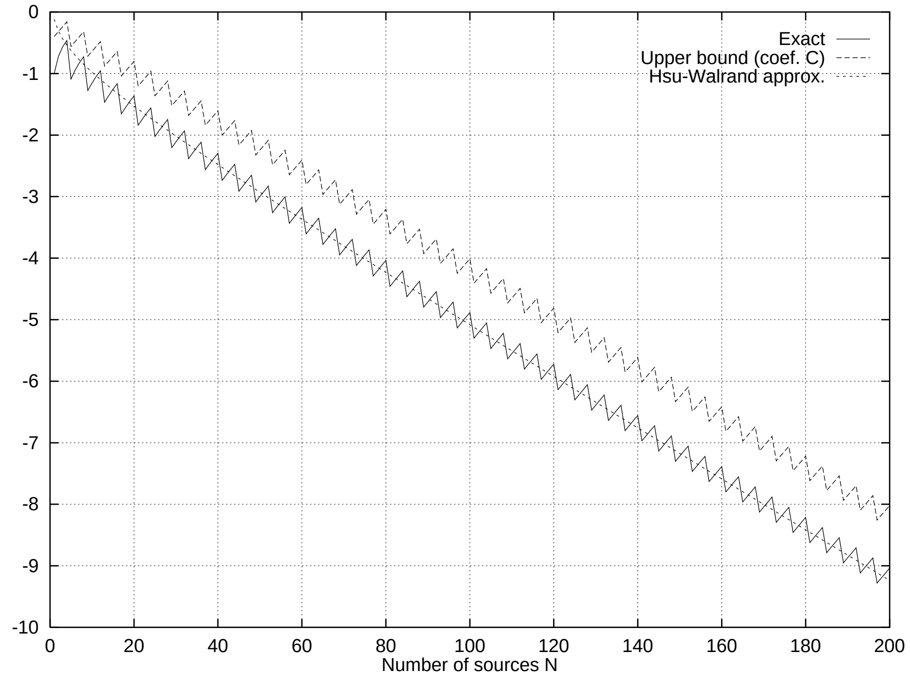

Figure 8: Plot of $\log_{10}(\Phi(N))$, with load=0.4 and $(1/\lambda, 1/\mu, r) = (45.0, 5.0, 1.0)$

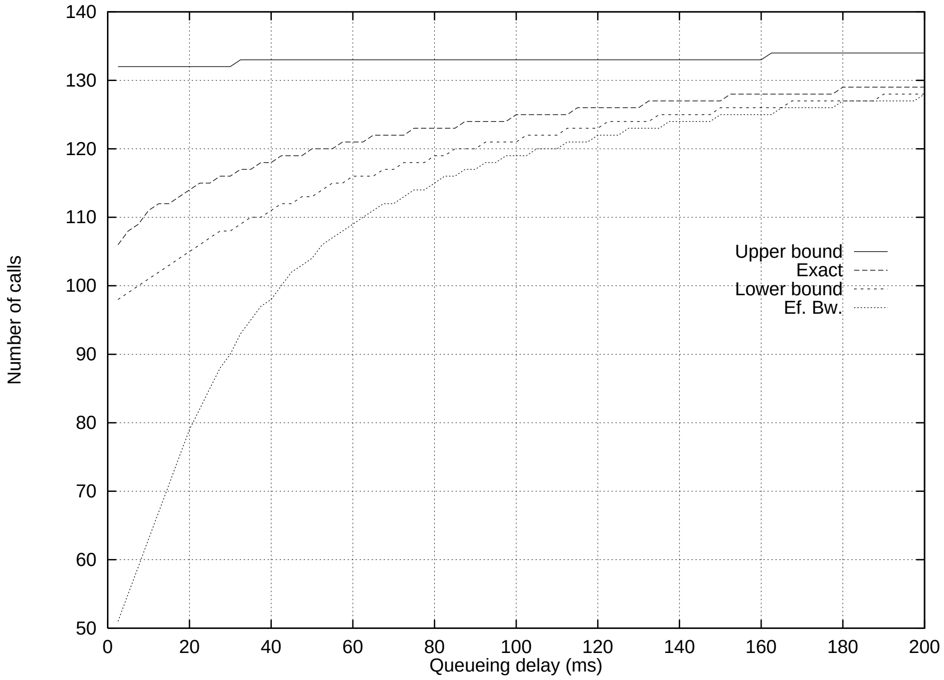

Figure 9: Number of admissible calls, with criterion $q = 1\%$.

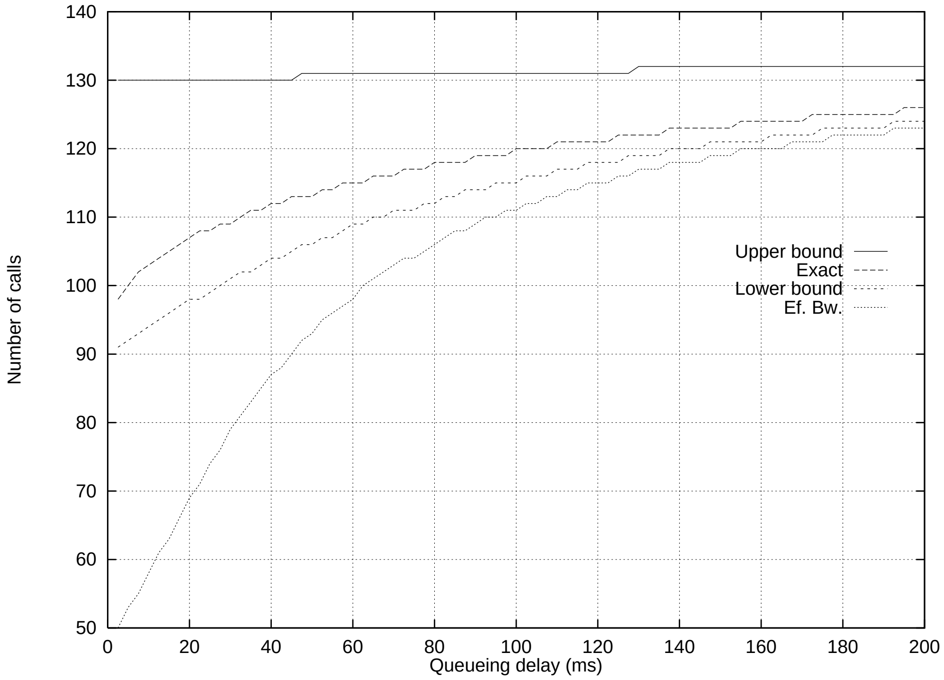

Figure 10: Number of admissible calls, with criterion $q = 0.1\%$.

- Number N_{EB}: this is the number determined by the effective bandwidth mechanism, and is simply equal to $\lfloor c/a(\theta_0) \rfloor$, with $\theta_0 = -\log(q)/b$.
- Lower bound N_i: this number is obtained by using the upper bound $C \exp(-\theta^\star x)$ on $\mathrm{P}(X > x)$. If the number of sources is such that the upper bound is less than q, then the QoS criterion is satisfied. Thus we have $N_i \leq N_e$, where $N_i = \max\{N \;/\; C \exp(-\theta^\star b) \leq q\}$, the coefficient C depending on N.
- Upper bound N_s: similarly, N_s is calculated by using the lower bound on $\mathrm{P}(X > b)$ with coefficient B.

Figures 9 and 10 compare the previous numbers as a function of b, with q =1% and q =0.1%. We observe that the algorithm based purely on an effective bandwidth approach (number N_{EB}) yields poor results, especially for small b, by accepting approximately only one half of the possible connections.

We also note that the number N_s, upper bound on the maximum number of admissible connections, is almost equal to the number corresponding to mean rate allocation, independently of b. Thus, the number N_s provides only trivial information, which is due to the poor quality of coefficient B for a large number of sources. However, the number N_i, lower bound on the maximum number of admissible connections, deviates from the exact number by no more than about 10% in both examples.

5 Conclusion

The multiplexing of Markovian on/off sources has received a lot of attention in the recent past, with most of the results concerning symmetrical systems. We have proposed exponential upper and lower bounds on the baklog distribution which are easily computed, and which hold for any number of different traffic classes. We have conducted numerical experiments to test the validity of the bounds, and have compared them with other authors' results whenever possible. When considering a symmetrical system, we retrieve some previously well known formulas as a special case. We have argued that in large systems, our lower bound may greatly underestimate the exact value, whereas the upper bound, as observed in the symmetrical case, is presumably reasonably close to the exact result for small buffers.

References

[1] D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Syst. Tech. J.*, 61(8):1871–1894, 1982.

[2] D. Artiges. *Contrôle et Évaluation de Réseaux de Télécommunications.* PhD thesis, Université de Nice Sophia Antipolis, Nice, France, February 1996. (http://www.inria.fr/mistral/personnel/Damien.Artiges/these.html).

[3] S. Asmussen and T. Roslki. Risk theory in a periodic environment: the Cramer-Lundberg approximation and Lundberg's inequality. *Mathematics of Opns. Res.*, 2:410–433, 1994.

[4] B. Bensaou, J. Guibert, J. W. Roberts, and A. Simonian. Performance of an ATM multiplexer queue in the fluid approximation using the Beneš approach. *Ann. Operat. Res.*, 49:137–160, 1994.

[5] E. Buffet and N. G. Duffield. Exponential upper bounds via martingales for multiplexers with Markovian arrivals. *J. Appl. Prob.*, 31:1049–1060, 1994.

[6] C. S. Chang. Stability, queue length and delay, Part II: Stochastic queueing networks. In *31st IEEE Conference on Decision and Control*, 1992.

[7] C. S. Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. Aut. Contr.*, 39(5):913–931, May 1994.

[8] C. Courcoubetis and R. Weber. Buffer overflow asynptotics for a switch handling many traffic sources. To appear in J. A. P., 1996.

[9] R. L. Cruz. A calculus for network delay, part I: Network elements in isolation. *IEEE Trans. Inf. Theory*, 37:114–131, January 1991.

[10] R. L. Cruz. A calculus for network delay, part II: Network analysis. *IEEE Trans. Inf. Theory*, 37:132–141, January 1991.

[11] G. de Veciana, G. Kesidis, and J. Walrand. Resource management in wide-area ATM networks using effective bandwidths. *IEEE J. Select. Areas Commun.*, 13:1081–1090, August 1995.

[12] N. G. Duffield. Exponential bounds for queues with Markovian arrivals. *Queueing Systems*, 17:413–430, 1994.

[13] A. Elwalid, D. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss. Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing. *IEEE J. Select. Areas Com.*, 13(6):1004–1016, August 1995.

[14] A. I. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. on Networking*, 1:329–343, June 1993.

[15] W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149–171, 1992.

[16] R. S. Garfinkel and G. L. Nemhauser. *Integer Programming.* John Wiley & Sons, 1972.

[17] R. J. Gibbens and P. J. Hunt. Effective bandwidth for the multi-type UAS channel. *Queueing Systems*, 9:17–27, 1991.

[18] R. J. Gibbens, F. P. Kelly, and P. B. Key. A decision-theoretic approach to call admission control in ATM networks. *IEEE J. Select. Areas Com.*, 13(6), August 1995.

[19] P. W. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Studies in Appl. Prob.*, pages 131–156, 1994.

[20] A. Graham. *Kronecker Products and Matrix Calculus with Applications.* Chischester: Ellis Horwood, 1981.

[21] R. Guérin, H. Ahmadi, and M. Naghshineh. Equivalent capacity and its application to bandwidth allocation in high-speed network. *IEEE J. Select. Areas Commun.*, 9:968–981, 1991.

[22] I. Hsu and J. Walrand. Admission control for ATM networks. In *IMA Workshop Stochastic Networks*, 1994.

[23] J. Y. Hui. Resource allocation for broadband networks. *IEEE J. Select. Areas Com.*, 6:1598–1608, 1988.

[24] F. P. Kelly. Effective bandwidths at multi-class queues. *Queueing Systems*, 9:5–16, 1991.

[25] G. Kesidis, J. Walrand, and C. S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Trans. Networking*, 1(4):424–428, August 1993.

[26] J. F. C. Kingman. Inequalities in the theory of queues. *J. Roy. Stat. Soc.*, 32:102–110, 1970. Series B.

[27] L. Kosten. Stochastic theory of a multi-entry buffer. Technical report, Unversity of Delft, The Netherlands, 1974.

[28] Z. Liu, P. Nain, and D. Towsley. Exponential bounds with an application to call admission. Technical Report CMPSCI 94-63, University of Massachusetts, October 1994. Submitted for publication. Abridged version in CDC'94.

[29] I. Norros, J. W. Roberts, A. Simonian, and J. Virtamo. The superposition of variable bit rate sources in ATM multiplexers. *IEEE J. Select. Areas Com.*, 9(3), April 1991.

[30] A. Simonian and J. Guibert. Large deviations approximation for fluid queues fed by a large number of on/off sources. In J. Labetoulle and J. W. Roberts, editors, *ITC 14*, pages 1013–1022. Elsevier, 1994.

[31] T. E. Stern and A. I. Elwalid. Analysis of separable markov-modulated models for information handling systems. *Adv. Appl. Prob.*, 1991.

[32] D. E. Stewart. *Meschach: Matrix Computations in C.* Centre for Mathematics and its Applications, School of Mathematical Sciences, Australian National University, Canberra, 1994. CMA Proceedings #32.

[33] D. N. C. Tse, R. G. Gallager, and J. N. Tsitsiklis. Statistical multiplexing of multiple time-scale Markov streams. *IEEE J. Select. Areas Com.*, 13(6), August 1995.

[34] A. Weiss. A new technique for analyzing large traffic systems. *Adv. Appl. Prob.*, 18:506–532, 1986.

[35] W. Whitt. Tail probabilities with statistical mutiplexing and effective bandwidths in multi-class queues. *Telecommunication Systems*, 2:71–107, 1993.

[36] O. Yaron and M. Sidi. Performance and stability of communication networks via robust exponential bounds. *IEEE/ACM Trans. Networking*, 1:372–385, 1993.