EFFECTIVE BANDWIDTHS FOR MULTICLASS MARKOV FLUIDS AND OTHER ATM SOURCES

by

G. Kesidis and J. Walrand

Memorandum No. UCB/ERL M92/40

27 April 1992

EFFECTIVE BANDWIDTHS FOR MULTICLASS MARKOV FLUIDS AND OTHER ATM SOURCES

by

G. Kesidis and J. Walrand

Memorandum No. UCB/ERL M92/40

27 April 1992

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

EFFECTIVE BANDWIDTHS FOR MULTICLASS MARKOV FLUIDS AND OTHER ATM SOURCES

by

G. Kesidis and J. Walrand

Memorandum No. UCB/ERL M92/40

27 April 1992

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources*

G. Kesidis[1,2] J. Walrand[1]

1. EECS Dept., University of California, Berkeley, CA94720.

2. Electrical Eng. Dept., University of Waterloo, Waterloo, ON, Canada, N2L 3G1.

*Supported by: NSERC of Canada, Pacific Bell, Micro Grant of the State of California, ARO Grant Number DAAL03-89-K-0128

Abstract

We show the existence of effective bandwidths for multiclass Markov fluids and other types of sources that are used to model ATM traffic. More precisely, we show that when such sources share a buffer with deterministic service rate a constraint on the tail of the buffer occupancy distribution is a linear constraint on the number of sources. That is, for a small loss probability one can assume that each source transmits at a fixed rate called its effective bandwidth. When traffic parameters are known, effective bandwidths can be calculated and may be used to obtain a circuit-switched style call acceptance and routing algorithm for ATM networks. The important feature of the effective bandwidth of a source is that it is a characteristic of that source, the buffer size and the acceptable loss probability. Thus, the effective bandwidth of a source does not depend on the number of sources sharing the buffer nor on the model parameters of other types of sources sharing the buffer.

1 Introduction

Effective bandwidths have been discovered for certain traffic models and certain performance criteria (see [1], [2], [3]). For example, consider a buffer of infinite size with service rate c cells/s. Let X be the number of cells in the buffer found by a typical arriving cell. Suppose that

$$\mathrm{P}\{X > B\} \;\leq\; \mathrm{e}^{-B\delta} \tag{1}$$

must be satisfied (the performance criterion). Suppose further that there are N_j independent on-off Markov fluids [4] of type j ($j = 1, 2, ..., K$) sharing the buffer. Gibbens and Hunt in [2] find constants α_i that depend only on the parameters of a type i source and δ, such that the constraint (1) holds for $B\delta >> 1$ if and only if

$$\sum_{j=1}^{K} N_j \alpha_j \;\leq\; c.$$

We call α_i the *effective bandwidth* of an on-off Markov fluid of type i.

In general, effective bandwidths depend on both the traffic/buffer models and the performance criterion. Kelly [1] finds effective bandwidths for GI/G/1 queues under (1) and for M/G/1 queues with the performance criterion taken to be the buffer utilization (fraction of time $X \neq 0$) or mean workload ($EX < B$). Courcoubetis and Walrand [3] find effective bandwidths for stationary Gaussian sources under (1). The open question answered in this paper is the existence of effective bandwidths for more general source

models under (1).

We start by heuristically deriving an expression for $\mathrm{P}\{X > B\}$ for general source models. Consider an infinite buffer with service rate c shared by N_i sources of type i, $i = 1, ..., K$. All the sources are assumed independent. For all M_i greater than the average rate of cells produced by a source of type i, assume that the probability that a source of type i produces $M_i T$ cells over a period of time of length T is approximately $\exp(-TH_i(M_i))$ where H_i is strictly convex and non-negative (this assumption follows from large deviations as we will see later on). By independence, the probability that, for $j = 1, ..., N_i$, the j^{th} source of type i produces $\mu_j T$ cells over time T is about

$$\exp\left(-T\sum_{j=1}^{N_i} H_i(\mu_j)\right).$$

Consequently, the probability that all sources of type i produce a total of $N_i M_i T$ cells over large time T is about

$$\sum_{\mu:\sum\mu_j=N_iM_i} \exp\left(-T\sum_{j=1}^{N_i} H_i(\mu_j)\right)$$

where $\mu = (\mu_1, ..., \mu_{N_i})$. Indeed, each choice of μ such that $\sum \mu_j = N_i M_i$ is one particular way for $N_i M_i T$ cells to get produced. This sum of exponentials can be approximated by the largest term (originally an argument of Laplace):

$$\sum_{\mu:\sum\mu_j=N_iM_i} \exp\left(-T\sum_{j=1}^{N_i} H_i(\mu_j)\right) \approx \exp\left(-\inf_{\mu:\sum\mu_j=N_iM_i} T\sum_{j=1}^{N_i} H_i(\mu_j)\right)$$

$$= \exp\left(-TN_iH_i(M_i)\right)$$

where the last equality is due to the convexity of H_i. Therefore, by independence, the probability that, for $i = 1, ..., K$, the sources of type i produce N_iM_iT over time T is about

$$\exp\left(-T\sum_{i=1}^{K} N_iH_i(M_i)\right).$$

Thus, the probability that, starting from an empty buffer, the sources of type i produce cells at rate N_iM_i until the buffer occupancy exceeds B is

$$\exp\left(-B\frac{\sum N_iH_i(M_i)}{\sum N_iM_i - c}\right).$$

Indeed $T = B/(\sum N_iM_i - c)$ is the time the buffer occupancy takes to reach B when the aggregate cell arrival rate is $\sum N_iM_i$. By the argument of Laplace, the probability that the buffer occupancy, starting from empty, reaches B before it returns to empty is about

$$\exp\left(-B\inf_{\sum N_iM_i > c}\frac{\sum N_iH_i(M_i)}{\sum N_iM_i - c}\right) \approx \mathrm{P}\{X > B\}. \tag{2}$$

This paper is organized as follows. In section 2, we show the existence of effective bandwidths in the multiclass case if the H_i satisfy certain conditions. In section 3, we give expressions for the H_i for Markov fluids, Markov-modulated Poisson processes, and discrete-time ergodic stationary sources. Finally, conclusions are drawn in section 4.

2 General Effective Bandwidths

We now show the existence of effective bandwidths. First, some assumptions on the H_i are made, then effective bandwidths are defined by considering the single source case, and finally the multiclass case is considered.

Consider an infinite buffer with deterministic service rate c cells/s, shared by N^i independent sources of type i, $i = 1, ..., K$. Denote by $[\cdot, \cdot]$ the scalar product. Let $\Gamma_i \in (0, \infty]$ (respectively $\gamma_i \in [0, \infty)$) denote the maximum (respectively minimum) possible cell arrival rate of a type i source. Let $\overline{\gamma}_i \in (0, \infty)$ be the average arrival rate of a type i source. We assume that

$$N \in \mathbf{C} := \{N \in \mathbf{R}_+^K : [N, \Gamma] > c \text{ and } [N, \overline{\gamma}] < c\}$$

where $\overline{\gamma} = (\overline{\gamma}_1, ..., \overline{\gamma}_K)$ and $\Gamma := (\Gamma_1, ..., \Gamma_K)$. Let $M = (M_1, ..., M_K)$.

By equation (2) we expect that

$$\mathrm{P}\{X > B\} = \exp(-BI(N, c) + o(B)),$$

where

$$I(N, c) = \inf_{M \in \mathbf{A}(N,c)} \frac{\sum_{i=1}^K N_i H_i(M_i)}{[N, M] - c}, \tag{3}$$

and $\mathbf{A}(N, c) := \{M \in \mathbf{R}_+^K : \gamma_i < M_i < \Gamma_i \ \forall i \text{ and } [N, M] > c\}$. Thus, the constraint

(1) is equivalent to

$$I(N,c) \geq \delta.$$

We make the following assumptions about H_i over the interval (γ_i, Γ_i) for all i:

[H.0] $\gamma_i < \overline{\gamma}_i < \Gamma_i$;

[H.1] H_i is strictly convex and $H_i \in \mathbf{C}^1(\gamma_i, \Gamma_i)$;

[H.2] $\infty > H_i(M_i) \geq 0$ with $H_i(M_i) = 0 \Leftrightarrow M_i = \overline{\gamma}_i$;

[H.3] $H_i'(\Gamma_i-) = \infty$ even when $\Gamma_i < \infty$.

As we will see later, assumptions **[H.1]** and **[H.2]** are an immediate consequence of the theory of large deviations. Assumption **[H.3]** will also be justified later.

We will need the following technical result.

Lemma 1: For all $a \in (\overline{\gamma}_i, \Gamma_i)$,

$$\lim_{M_i \to \Gamma_i} \frac{H_i(M_i)}{H_i'(M_i)(M_i - a)} < 1.$$

Proof: Fix $a \in (\overline{\gamma}_i, \Gamma_i)$. Note that if $\Gamma_i = \infty$, then by convexity

$$1 > \lim_{M_i \to \infty} \frac{H_i(M_i)}{H_i'(M_i)(M_i - \overline{\gamma}_i)} = \lim_{M_i \to \infty} \frac{H_i(M_i)}{H_i'(M_i)(M_i - a)}.$$

If $\Gamma_i < \infty$ and $H(\Gamma_i-) < \infty$ then, by **[H.3]**,

$$\lim_{M_i \to \Gamma_i} \frac{H_i(M_i)}{H_i'(M_i)(M_i - a)} = 0.$$

Finally, if $\Gamma_i < \infty$ and $H(\Gamma_i-) = \infty$, define $\phi_i(y) := H_i(\Gamma_i - y^{-1})$ for $y \in ((\Gamma_i - \overline{\gamma}_i)^{-1}, \infty)$. Note that ϕ is strictly convex and

$$\begin{aligned}
\lim_{M_i \to \Gamma_i} \frac{H_i(M_i)}{H_i'(M_i)(M_i - a)} &= \frac{1}{\Gamma_i - a} \lim_{y \to \infty} \frac{\phi_i(y)}{\phi_i'(y) y^2} \\
&< \frac{1}{\Gamma_i - a} \lim_{y \to \infty} \frac{\phi_i'(y)(y - (\Gamma_i - \overline{\gamma}_i)^{-1})}{\phi_i'(y) y^2} \\
&= 0. \quad \spadesuit
\end{aligned}$$

2.1 Single Source Case

Consider an infinite buffer with a single source of type i. For $\delta > 0$, define $\alpha_i(\delta)$ to be the value of a such that

$$I_i(a) := \inf_{M_i \in \mathbf{A}_i(a)} \frac{H_i(M_i)}{M_i - a} = \delta$$

where $\mathbf{A}_i(a) := \{M_i : a < M_i < \Gamma_i\}$. Thus, $\alpha_i(\delta) = I_i^{-1}(\delta)$ can be interpreted as the rate at which to serve a single source of type i so that the constraint (1) is satisfied. We call α_i is called the *effective bandwidth* of the type i traffic. At this point we show that α_i is well defined.

Lemma 2: α_i is a continuous, increasing function on $(0, \infty)$ with range $(\overline{\gamma}_i, \Gamma_i)$.

Proof: For $a \in (\overline{\gamma}_i, \Gamma_i)$, define

$$f_i(a, M_i) \;:=\; \frac{H_i(M_i)}{M_i - a},$$

so that $I_i(a) \;:=\; \inf_{M_i \in \mathbf{A}_i(a)} f_i(a, M_i)$. To evaluate I_i note that,

$$\frac{\partial f_i}{\partial M_i}(a, \hat{M}_i) = 0 \;\Leftrightarrow\; g_i(a, \hat{M}_i) = 0$$

where

$$g_i(a, M_i) \;:=\; H_i'(M_i)(M_i - a) - H_i(M_i). \tag{4}$$

By assumptions **[H.0]**–**[H.2]**, $g_i(a, a) < 0$ and $g_i(a, \cdot)$ is increasing on $[a, \Gamma_i)$. By lemma 1, $g_i(a, \Gamma_i -) > 0$. See Figure 1. Thus, $f_i(a, \cdot)$ is unimodal on $[a, \Gamma_i)$ and there exists a unique $\hat{M}_i(a) \in (a, \Gamma_i)$ such that $g_i(a, \hat{M}_i(a)) = 0$ and

$$I_i(a) \;=\; f_i(a, \hat{M}_i(a)) \;=\; H_i'(\hat{M}_i(a)).$$

Note that $\hat{M}_i(a)$ is continuous by **[H.1]**, which implies that I_i is continuous as well.

We will now show that $I_i(\cdot)$ is increasing on $(\overline{\gamma}_i, \Gamma_i)$. Clearly $\hat{M}_i(\Gamma_i -) = \Gamma_i$. Also, assumptions **[H.0]**–**[H.2]** imply that $H_i'(M_i) > H_i(M_i)/(M_i - \overline{\gamma}_i)$ for all $M_i \in (\overline{\gamma}_i, \Gamma_i)$. Therefore $\hat{M}_i(\overline{\gamma}_i +) = \overline{\gamma}_i$. Finally, by differentiating $g_i(a, \hat{M}_i(a)) \equiv 0$ implicitly, we find

that

$$\hat{M}_i'(a) = \frac{H_i'(\hat{M}_i(a))}{H_i''(\hat{M}_i(a))(\hat{M}_i(a) - a)} > 0$$

for all $a \in (\overline{\gamma}_i, \Gamma_i)$. Therefore $I_i(a) = H_i'(\hat{M}_i(a))$ is an increasing function on the interval $(\overline{\gamma}_i, \Gamma_i)$ with $I_i(\overline{\gamma}_i+) = 0$ and $I_i(\Gamma_i-) = \infty$ as desired. ♠

Assumption **[H.3]** is now physically justified in the following way. Given an infinite buffer with service rate a cells/s and a single type i source, the probability that the buffer occupancy, starting from empty, reaches B before it returns to empty is about $\exp(-BI_i(a))$, for all sufficiently large B. Therefore, as $a \to \Gamma_i < \infty$ we expect that $\exp(-BI_i(a)) \to 0$, for all sufficiently large B. Thus, as $a \to \Gamma_i$, $I_i(a) \to \infty$. The statement of assumption **[H.3]** now follows from the fact that $I_i(a) = H_i'(\hat{M}_i(a))$ and that $\hat{M}_i(a) \to \Gamma_i$ as $a \to \infty$. When H_i does not have a continuous derivative ([**H.2**]), an identical assumption to **[H.3]** can be made on the generalized gradient of H_i instead of H_i'.

2.2 Multiclass Case

In this section we evaluate $I(N, c)$ of equation (3) for $N \in \mathbf{C}$ and prove the "effective bandwidth theorem".

For all $\delta > 0$, define the set

$$\mathbf{B}(\delta) = \left\{ N \in \mathbf{R}_+^K : \sum N_i \alpha_i(\delta) = c \right\}.$$

Note that, by lemma 2, $\mathbf{B}(\delta) \subset \mathbf{C}$ for all $\delta > 0$. Take $N \in \mathbf{B}(\delta)$ and recall that, for all i,

$$I_i(\alpha_i(\delta)) \;=\; \delta \;=\; H_i'(\hat{M}_i(\alpha_i(\delta))). \tag{5}$$

Recall that

$$I(N,c) \;=\; \inf_{M \in \mathbf{A}(N,c)} F(N,c,M)$$

where

$$F(N,c,M) \;:=\; \frac{\sum N_i H_i(M_i)}{[N,M]-c}.$$

Lemma 3: For all $\delta > 0$, $N \in \mathbf{B}(\delta)$ implies $I(N,c) = \delta$.

Proof: By differentiating we find that

$$\begin{aligned} \frac{\partial F}{\partial M_i}(N,c,M) = 0 \;&\Leftrightarrow\; H_i'(M_i)([N,M]-c) - \sum N_j H_j(M_j) = 0 \\ &\Leftrightarrow\; H_i'(M_i) = \frac{\sum N_j H_j(M_j)}{[N,M]-c} \;=\; F(N,c,M). \end{aligned}$$

Therefore,

$$\nabla_M F(N,c,M) = 0 \;\Rightarrow\; \sum N_i \left\{ H_i'(M_i)(M_i - \alpha_i(\delta)) - H_i(M_i) \right\} = 0.$$

Equivalently (definition (4)),

$$\nabla_M F(N,c,M) = 0 \;\Rightarrow\; \sum N_i g_i(\alpha_i(\delta), M_i) = 0. \tag{6}$$

Therefore, if $M_i = \hat{M}(\alpha_i(\delta))$ for all i, then $\nabla_M F(N,c,M) = 0$ (see equation (5)).

So we have shown that for any $\delta > 0$ and for all $N \in \mathbf{B}(\delta)$,

$$\nabla F(\hat{M}(\alpha(\delta))) = 0 \quad \text{and} \quad F(\hat{M}(\alpha(\delta))) = H_i'(\hat{M}_i(\alpha_i(\delta))) = \delta$$

where $\hat{M}(\alpha(\delta)) = (\hat{M}_1(\alpha_1(\delta)), ..., \hat{M}_K(\alpha_K(\delta)))$.

What remains to show is that $\hat{M}(\alpha(\delta)) \in \mathbf{A}^o(N,c)$ is the unique global minimizer of $F(N,c,\cdot)$ on $\mathbf{A}(N,c)$. Assumptions **[H.2]** and **[H.3]** imply that the minimum of F cannot be on the boundary of $\mathbf{A}$. For a proof by contradiction, assume that there is an $S \in \mathbf{A}^o(N,c)$, $S \neq \hat{M}(\alpha(\delta))$, such that $I(N,c) = F(S) = H_i'(S_i) < \delta$ for all i. Recall that H_i' is an increasing function and $H_i'(\hat{M}_i(\alpha_i(\delta))) = \delta$. Therefore, $I < \delta$ implies $S_i < \hat{M}_i(\alpha_i(\delta))$ for all i. Also if $S_i \in (\overline{\gamma}_i, \hat{M}_i(\alpha_i(\delta)))$ for all i, then $g_i(\alpha_i(\delta)), S_i) < 0$ for all i (see Figure 1). Thus, the necessary condition (6) is violated. Therefore, there exists an i^* such that $S_{i^*} < \overline{\gamma}_{i^*} \Rightarrow 0 > H_{i^*}'(S_{i^*}) = F(S) = I(N,c)$; but $F \geq 0$ on $\mathbf{A}$ so we have a contradiction. ♠

The following theorem justifies the interpretation of the α_i as effective bandwidths in the multiclass case.

Theorem: For any $\delta > 0$ and $N \in \mathbf{C}$,

$$I(N,c) \geq \delta \;\Leftrightarrow\; \sum N_i \alpha_i(\delta) \leq c.$$

Proof: We first show that $N \in \mathbf{B}(\delta)$ if and only if $I(N,c) = \delta$. By lemma 3, it suffices to show that $N \in \mathbf{C} \cap \overline{\mathbf{B}(\delta)}$ implies $I(N,c) \neq \delta$. So assume $N \in \mathbf{C} \cap \overline{\mathbf{B}(\delta)}$. Recall that $[N, \alpha(\cdot)]$ is continuous and $N \in \mathbf{C}$ implies that $[N, \alpha(0+)] = [N, \overline{\gamma}] < c$ and $[N, \alpha(\infty)] = [N, \Gamma] > c$. Therefore, there exists an $\varepsilon \neq \delta$, $\varepsilon > 0$, such that $[N, \alpha(\varepsilon)] = c$ (i.e., $N \in \mathbf{B}(\varepsilon)$) which implies that $I(N,c) = \varepsilon$, by lemma 3, as desired.

Thus for $N \in \mathbf{C}$, $I(N,c) = \delta$ if and only if $[N, \alpha(\delta)] = \sum N_i \alpha_i(\delta) = c$. Since the α_i are all increasing, $[N, \alpha(\cdot)]$ is increasing. Therefore if $I(N,c) = \varepsilon > \delta$ then $c = [N, \alpha(\varepsilon)] > [N, \alpha(\delta)]$. Conversely, if $[N, \alpha(\delta)] < c$, there exists an $\varepsilon > \delta$ such that $c = [N, \alpha(\varepsilon)]$ which implies that $I(N,c) = \varepsilon$. Thus

$$\begin{aligned} \{N \in \mathbf{C} \,:\, I(N,c) \geq \delta\} &= \bigcup_{\varepsilon \geq \delta} \mathbf{B}(\varepsilon) \\ &= \{N \in \mathbf{C} \,:\, [N, \alpha(\delta)] \leq c\}. \;\spadesuit \end{aligned}$$

3 Models of ATM Buffer Sources

We now consider three models of buffer sources used to characterize bursty ATM traffic. We will interpret the assumptions **[H.1]**-**[H.3]** for each model considered.

3.1 Markov Fluids

A source is called a Markov fluid if its time-derivative is a continuous-time Markov chain on a finite state space. If the arrival process to a buffer with deterministic service rate *is* a superposition of independent Markov fluids, then the buffer occupancy has piecewise-linear trajectories with random slopes.

For each type i of Markov fluid ($i = 1, ..., K$), we let $\Lambda^i = (\Lambda^i_1, ..., \Lambda^i_{m_i})$, be the state space and Q^i be the transition rate matrix of its Markov time-derivative. We assume $\Lambda^i_j < \Lambda^i_{j+1} < \infty$ for all i, j. Therefore, in the notation of section 2, $\gamma_i = \Lambda^i_1$, $\Gamma_i = \Lambda^i_{m_i}$, and $\overline{\gamma}_i := [\pi^i, \Lambda^i]$ where π^i is the invariant of Q^i: $\pi^i Q^i = 0$.

To define H_i, let $J^c_{Q^i}$ be the large deviations action functional for the empirical distribution of a continuous-time Markov chain with transition rate matrix Q^i (see the Appendix). Take

$$H_i(M_i) \;:=\; \inf_{[\mu,\Lambda^i]=M_i} J^c_{Q^i}(\mu) \tag{7}$$

where the infimum is taken over the space Σ_{m_i} of distributions on Λ^i. Note that the strict convexity of $J^c_{Q^i}$ on Σ_{m_i} implies that H_i is strictly convex on $(\gamma_i, \Gamma_i) = (\Lambda^i_1, \Lambda^i_{m_i})$.

3.1.1 Two-State Markov Fluids Example

For example, if the Markov fluids are all of the two state ($m_i = 2$) type, then (see equation (7))

$$H_i(M_i) \;=\; \frac{1}{\Lambda_2^i - \Lambda_1^i}\left(\sqrt{q_1^i(\Lambda_2^i - M_i)} - \sqrt{q_2^i(M_i - \Lambda_1^i)}\right)^2$$

where $q_1^i = Q_{1,2}^i$ and $q_2^i = Q_{2,1}^i$. Note that assumptions **[H.1]**-**[H.3]** are satisfied. The single class case gives [4]:

$$I_i(a) \;=\; \frac{(q_1^i + q_2^i)(a - \overline{\gamma}_i)}{(a - \Lambda_1^i)(\Lambda_2^i - a)}.$$

By direct calculation, the effective bandwidths are

$$\alpha_i(\delta) \;=\; I_i^{-1}(\delta) \;=\; \frac{1}{2}\left(-a_i(\delta) + \sqrt{a_i^2(\delta) - 4b_i(\delta)}\right)$$

where

$$a_i(\delta) \;=\; \frac{q_1^i + q_2^i}{\delta} - \Lambda_2^i - \Lambda_1^i \quad \text{and} \quad b_i(\delta) \;=\; \Lambda_2^i\Lambda_1^i - \frac{q_1^i\Lambda_2^i + q_2^i\Lambda_1^i}{\delta}$$

Gibbens and Hunt [2] take $\Lambda_1^i = 0$ for all i ("on-off" Markov fluids).

3.2 Markov-Modulated Poisson Process

A source to a buffer is called a Markov-modulated Poisson process (MMPP) if the cell arrivals are Poisson with intensity λ where λ is a continuous-time Markov chain (i.e., an MMPP is a Poisson process with Markov intensity). Again, we have a state space Λ^i and the transition rate matrix Q^i for each source of type i with $\overline{\gamma}_i := [\pi^i, \Lambda^i]$ as in the Markov fluid case. However, $\gamma_i = 0$ and $\Gamma_i = \infty$ for an MMPP source.

To define H_i for MMPPs, first let Y_a be a random variable with Poisson distribution and mean a^{-1}. Define (see Cramér's theorem in [5]):

$$L(a,x) \;:=\; \sup_{y\in\mathbf{R}}\left\{xy \;-\; \log \mathrm{E}e^{yY_a}\right\} \;=\; x\log\frac{x}{a} \;+a-x.$$

The contraction mapping principle [6] motivates us to use

$$H_i(M_i) \;:=\; \inf_{\mu\in\Sigma_{m_i}} \left\{J^c_{Q^i}(\mu) + L([\mu,\Lambda^i], M_i)\right\}.$$

Note that, since $L(\cdot,\cdot)$ is a strictly convex function on $\mathbf{R}^2_+$, the strict convexity of J^c_Q implies that H_i is strictly convex on $(\gamma_i, \Gamma_i) = (0,\infty)$.

3.3 Discrete-Time Sources

A discrete-time source is called a Markov chain if the *number* of cell arrivals of that source, at each point in discrete-time, is a Markov chain on a finite state space Λ^i. In this case, the H_i are found by using the Donsker-Varadhan action functional $J^d_{Q^i}$ for

discrete-time Markov chains (see the Appendix), instead of $J^{\epsilon}_{Q^i}$, in equation (7)). Q^i is the transition probability matrix of a type i source in this case.

More generally, consider a single source of type i where the number of cell arrivals at time n is Z_n. Assume Z_n is a stationary and ergodic process satisfying the conditions of the Gärtner-Ellis theorem [7]. That is, assume

$$h_i(y) \;:=\; \lim_{n\to\infty} \frac{1}{n} \log \mathrm{E} \exp\left((Z_1 + ... + Z_n)y\right)$$

exists and is finite for all real y, and that h_i is differentiable whenever $h_i < \infty$. We then take H_i to be the Legendre transform of h_i:

$$H_i(M_i) \;:=\; \sup_{y\in\mathbf{R}} \{yM_i - h_i(y)\}.$$

Also, $\gamma_i = \inf\{\gamma : \mathrm{P}\{Z < \gamma\} > 0\}$, $\Gamma_i = \sup\{\Gamma : \mathrm{P}\{Z > \Gamma\} > 0\}$, and $\overline{\gamma}_i = \mathrm{E}Z$. Clearly H_i is non-negative ($y = 0$), convex and lower semi-continuous. Also, $H_i(\overline{\gamma}_i) = 0$ ($y = 1$). So, what remains to check is that H_i is strictly convex, $\mathbf{C^1}$, and **[H.3]** for the existence of equivalent bandwidths.

Assuming H_i is $\mathbf{C^1}$, we can directly check **[H.3]** when the Z_i are i.i.d. First note that $H_i'(M_i) = h_i'^{-1}(M_i)$ where

$$M_i \;=\; h_i'(y) \;=\; \frac{\mathrm{E}Ze^{yZ}}{\exp h_i(y)}.$$

Letting p be the distribution of the Z_j on $[\gamma_i, \Gamma_i]$, we find that, as $M_i \to \Gamma_i$,

$$\frac{e^{yz}p(z)}{\exp h_i(y)} \quad \to \quad \delta_{\Gamma_i}(z)$$

(i.e., $y \to \infty$). Thus, $H_i'(M) \to \infty$ as $M_i \to \Gamma_i$.

4 Conclusions

We have shown the existence of effective bandwidths for Markov fluids and other models of ATM traffic. Given effective bandwidths, one can determine the *spare capacity* of a buffer at any time. For instance, say we want to determine if a call of type j can be accommodated (i.e., constraint (1) is preserved) in a buffer that is currently being used by N_i calls of type i, $i = 1, ..., K$. If $\alpha_j < c - [N, \alpha(\delta)]$ then the call can be accommodated, else it cannot. Note that these are asymptotic results that hold when the loss probability is small. The practical relevance of these results should be explored for specific parameter values.

References

[1] F. Kelly, "Effective bandwidths at multi-class queues," *preprint.*

[2] R. Gibbens and P. Hunt, "Effective bandwidths for multi-type UAS channel," *submitted to QUESTA.*

[3] C. Courcoubetis and J. Walrand, "Note on effective bandwidth of ATM traffic," *preprint.*

[4] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," *Bell Sys. Tech. J.*, vol. 61 No.8, pp. 1871–1894, 1982.

[5] J.-D. Deuschel and D. W. Stroock, *Large Deviations.* New York, NY: Academic Press, 1989.

[6] J. Lynch and J. Sethuraman, "Large deviations for processes with independent increments," *Annals of Probability*, vol. 15 No. 2, pp. 610–627, 1987.

[7] J. Bucklew, *Large Deviation Techniques in Decision, Simulation and Estimation.* New York, NY: John Wiley and Sons, Inc., 1990.

[8] G. Kesidis, "Estimation of cell loss in high speed digital networks," *Ph.D. Dissertation, EECS Dept, U.C. Berkeley*, 1992.

5 Appendix: Donsker-Varadhan Action Functionals

For completeness, we give the following expression for J_Q in continuous and discrete time.

In discrete time [7], Q is a transition probability matrix on a state space $\Lambda = (\Lambda_1, ..., \Lambda_m)$.

For $\mu \in \Sigma_m$,

$$J_Q^d(\mu) = \inf_{P\,:\,\mu P = \mu} G^d(P;Q)$$

where the infimum is taken over the space of transition probability matrices on Λ, G^d is the relative entropy rate between discrete-time Markov chains,

$$G^d(P;Q) = \sum_{i,j=1}^{m} \mu_i P_{i,j} \log \frac{P_{i,j}}{Q_{i,j}},$$

and μ is the invariant of P ($\mu P = \mu$) and $\log \frac{0}{0} := 0$.

In continuous time ([5], p.125-128), Q is a transition rate matrix on Λ. For $\mu \in \Sigma_m$,

$$J_Q^c(\mu) = \inf_{P \,:\, \mu P = 0} G^c(P;Q)$$

where the infimum is taken over the space of transition rate matrices on Λ, G^c is the relative entropy rate between continuous-time Markov chains [8],

$$G^c(P;Q) \;:=\; \sum_{i=1}^{m} \mu_i \sum_{j=1, j\neq i}^{m} \left(P_{i,j} \log \frac{P_{i,j}}{Q_{i,j}} + Q_{i,j} - P_{i,j} \right),$$

and μ is the invariant of P ($\mu P = 0$). This definition of J_Q^c is different but consistent–in the sense of the contraction mapping principle [6]–with that in [5] (see "level 2.5" large deviations in [8]).

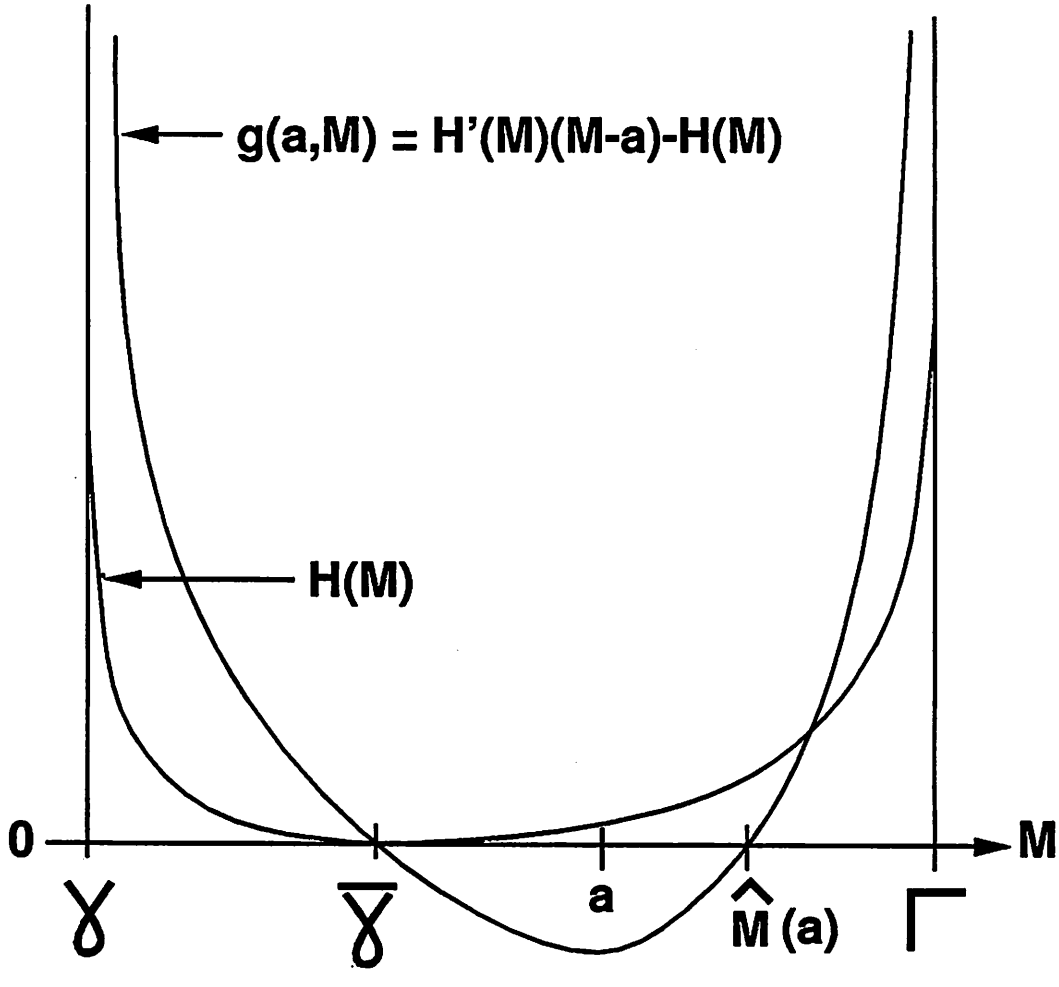

Figure 1: Plot of the function g