Optimal Operating Point in MIMO Channel for Delay-Sensitive and Bursty Traffic

Somsak Kittipiyakul and Tara Javidi
Department of Electrical and Computer Eng.
University of California, San Diego
La Jolla, CA 92093, USA
Email: skittipi@ucsd.edu and tara@ece.ucsd.edu

***Abstract*— We consider a system with a bursty and delay-sensitive data source to be transmitted over a constant-rate MIMO channel with no CSI information at the transmitter. Given the diversity-multiplexing tradeoff region of the MIMO channel, we find the optimal multiplexing rate that optimizes the end-to-end loss probability. Based on the effective bandwidth model of the source, we present an analytical tradeoff between the error probability over the MIMO channel and the probability of delay violation. We illustrate the optimal operating points for i.i.d. sources and Markov-modulated sources and show the relation between source burstiness, delay bound, and optimal multiplexing rate.**

I. Introduction

Multiple antennas are an important means to improve the performance of wireless systems. A MIMO system can provide two types of gains: diversity gain and multiplexing gain. Multiple antennas can be used to increase diversity to combat channel fading and thus to decrease the probability of error. Alternatively, multiple antennas can increase data rate by taking advantage of different fading over the antennas to multiplex independent data over these spatial channels. In [1], it is shown that diversity gain and multiplexing gain can be obtained simultaneously, but there is a tradeoff between the two gains: higher multiplexing gain comes at the price of lower diversity.

Our goal in this paper is to answer the question posed by Holliday and Goldsmith in [2]: ”given the diversity-multiplexing region, where should one choose to operate?”. To answer this question, they consider a system consisting of a source encoder concatenated with a MIMO channel encoder. Their goal is to determine the optimal operating point on the diversity-multiplexing region that minimizes the end-to-end distortion due to both the source encoder and channel decoding errors.

We answer the same question in a different context. We consider a system with a bursty and delay-sensitive data source, concatenated with an infinite buffer, followed by a constant rate MIMO channel encoder (see Figure 1). Due to the burstiness of the source, the arrival bits must be buffered for transmission over the fixed rate MIMO channel. Any bits out of the decoder which are delayed more than an acceptable threshold is considered obsolete by the receiving application. This is in addition to the error bits caused by the channel decoder. Our goal is to find the optimal operating point on the diversity-multiplexing region that minimizes the end-to-end bit loss probability due to both the delay bound violation and the MIMO channel decoding errors.

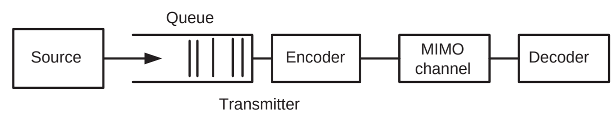


Fig. 1. Joint queuing and MIMO system.

Our work, in spirit, is related to [3]. In [3], the authors study the optimal (time-varying) encoding rate of a system consisting of a delay-sensitive constant rate source over a time-varying channel. The channel code is not assumed ideal. The channel encoder receives data from the queue at a time-varying rate and must encode the data at a rate matching the instantaneous channel quality. If the queue chooses a high instantaneous output rate, then the encoder must choose channel codes with large rate, thus clearing the queue quickly, but resulting in high decoding error probability. Their result shows that the exponent of the optimal end-to-end bit loss probability is the minimum of the delay bound violation exponent and the coding error exponent. As we will see, we arrive at similar results in the context of a time-varying source with MIMO encoder which assumes no knowledge of the MIMO channel and thus operates at a constant rate.

The paper is organized as follows. We begin in Section II with a background on the diversity-multiplexing region for MIMO channel and on the effective bandwidth. We formulate our problem in Section III and provide problem analysis in Section IV. The main result, the optimal operating point, is shown in Section V. In section VI, we apply the technique to three source models to illustrate their operating points. We conclude in Section VII.

II. Background

A. MIMO Channel Model

We use the same channel model as in [1]. We consider a wireless link with M transmit and N receive antennas. The fading coefficients h_{ij} that model the complex path gain from transmit antenna j to receive antenna i are i.i.d. complex circular symmetric Guassian with unit variance. The channel gain matrix $H = [h_{ij}] \in \mathcal{C}^{N\times M}$ is assumed to be known to the receiver but not at the transmitter. We assume that the channel remains constant over a block of T symbols, while each block is i.i.d. Therefore, in each block we can represent the channel as

$$Y = \sqrt{\frac{\rho}{M}} HX + W, \tag{1}$$

where $X \in \mathcal{C}^{M\times T}$ and $Y \in \mathcal{C}^{N\times T}$ are the transmitted and received signals, respectively; ρ is the average SNR at each receive antenna; The additive noise vector W is i.i.d. complex Guassian with unit variance.

As in [1], a family of codes $\{\mathcal{C}(\rho)\}$ of block length T, one codebook at each SNR ρ level, can be constructed. Define $R(\rho)$ as the number of bits per symbol for the codebook and $P_e(\rho)$ as the average probability of error for the codebook. A channel code scheme $\{\mathcal{C}(\rho)\}$ is said to achieve *multiplexing gain* r and *diversity gain* d if

$$\lim_{\rho\to\infty} \frac{R(\rho)}{\log \rho} = r \tag{2}$$

and

$$\lim_{\rho\to\infty} \frac{\log P_e(\rho)}{\log \rho} = -d. \tag{3}$$

Similar to [1], we will use the special symbol $\doteq$ to denote *exponentially equality*, i.e. $f(x) \doteq e^{bx}$ as a shorthand for $\lim_{x\to\infty} \frac{\log f(x)}{x} = b$.

For each r, define the optimal diversity gain $d^*(r)$ as the supremum of the diversity gain achieved by any scheme. The main result of [1] is summarized in the following statement.

Diversity-Multiplexing Tradeoff [1]: Assume the block length $T \geq M + N - 1$. Then the optimal tradeoff between diversity gain and multiplexing gain is a piecewise-linear function $d^*(r) = (M - r)(N - r)$, for $0 \leq r \leq \min(M, N)$, shown in Figure 2.

The diversity-multiplexing tradeoff is essentially the trade-off between the error probability and the data rate of a system, in the asymptotics of high SNR with fixed block length.

We assume without loss of generality that the rate of the codebook is $R(\rho) = rT \log \rho$ bits per block. Also, we assume that for any multiplexing gain r there is a codebook that achieves the optimal diversity gain $d^*(r)$.

B. Source Model and Effective Bandwidth

The source is modeled by a stochastic arrival stream $\{X_t\}$, where X_t indicates the number of bits arrived at time slot t. The arrived bits are queued at an infinite buffer and served with a fixed rate C in first-come-first-served manner. Since we are interested in real-time applications with strict delay bounds, it is of an interest to study the queue delay statistics and their dependence on the statistical characteristics of $\{X_t\}$. In particular, we are interested in the tail probability of the form $\mathcal{P}(Q \geq B)$, where Q is the steady-state queue length and B is the maximum acceptable bound. The key result from

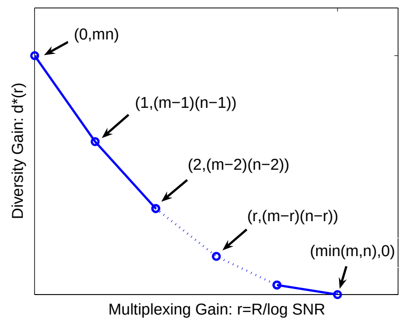


Fig. 2. Diversity-multiplexing tradeoff, $d^*(r)$ for general m, n, and $T \geq m + n - 1$.

effective bandwidth and large deviation literature (for example, [4] and [7]) is that, in the asymptotic regime for large B, the tail probability decays exponentially with B. More precisely,

$$\lim_{B\to\infty} \frac{1}{B} \log \mathcal{P}(Q \geq B) \leq -\delta \tag{4}$$

where the tail probability exponent δ is the solution to the following equation:

$$\alpha(\delta) = C. \tag{5}$$

The increasing function $\alpha(s)$, $s \geq 0$, is called *effective bandwidth* and is fully determined by the process $\{X_t\}$ (for a formal discussion on effective bandwidth see [4] and the references therein). [1] In this paper we consider the following two general source models:

1) I.I.D. Sources: Consider a source for which $\{X_t\}$ are i.i.d. random variables. The effective bandwidth is given by

$$\alpha(\delta) = \frac{\Lambda(\delta)}{\delta} \tag{6}$$

where $\Lambda(\delta) := \log E[e^{\delta X_t}]$ is the *log moment generating* function of X_t (for derivation of these results see [5]).

In Section VI, we will consider a compound Poisson process as an example of this type of the sources.

2) Markov-Modulated Sources: The arrival stream $\{X_t\}$ is modulated by a discrete-time, finite-state, irreducible, stationary Markov chain $\{H_t\}$ such that the distribution of X_t at time time t depends only on the source state H_t at time t. Given a realization of the chain $\{H_t\}$, the X_t's are independent. The source state H_t can be thought of as modeling the burstiness of the stream at time t. The Markov structure models the correlation in the arrival statistics over time [6]. It is shown that Markov-modulated source also has an effective bandwidth in the form of (6) but $\Lambda(\delta)$ is instead the *log spectral radius* function (see [6], [7], and [8]).[2] In Section VI, we will consider an on-off Poisson process as an example of Markov-modulated sources.

Note that Markov-modulated source (with multiple time scales) is often used to model MPEG video traffic [9].

[1]Note that the average and peak arrival rates are $\alpha(0)$ and $\alpha(\infty)$, respectively.

[2]In fact, the i.i.d. source is a degenerated case of Markov-modulated sources where the Markov chain is degenerated to only one state.

III. Problem Formulation

We begin by describing formally the system model. We consider time-slotted system[3] with X_t being the number of bits generated at the source into an infinite buffer in time slot t. The arrivals that are not immediately transmitted on the outgoing channel are queued up in the buffer. Without loss of generality, we assume a timeslot contains T symbols. Thus, a timeslot matches with the codebook block size. At SNR ρ and multiplexing rate r, the queue is served by a MIMO channel encoder at a fixed rate (bits per timeslot) of

$$C = rT \log \rho. \tag{7}$$

We consider a bursty and delay-sensitive application where any bits delayed more than D timeslots are considered obsolete. The source process is modeled as a source with an effective bandwidth of $\alpha(s)$, defined for all $s \geq 0$. Some examples of the source processes are given in Section VI. We assume that the average arrival rate $\alpha(0)$ is scaled with $\log \rho$, i.e. we define a constant $\lambda > 0$ as the following:

$$\lambda := \frac{\alpha(0)}{T \log \rho}. \tag{8}$$

Moreover, we assume that $\alpha(0) < C$. In another word, by (7) and (8), we assume the following condition on the multiplexing rate r:

$$r > \lambda. \tag{9}$$

There are two causes of performance loss in the system: Any obsolete bits out of the decoder are considered lost by the receiving application. In addition, error bits due to decoding in the MIMO channel are not retransmitted (i.e. no channel ARQ) and considered lost as well. Let P_q denote the probability of bit loss due to delay bound violation, P_e the probability of bit loss due to MIMO decoding errors, and P_t the end-to-end total bit loss probability that is perceived by the receiving application. By union bound, we have

$$P_t \leq P_q + P_e. \tag{10}$$

Intuitively we expect that P_q is decreasing on the multiplexing rate r because the higher rate the queue is served, the less time the bits spend waiting in the queue. On the other hand, from the diversity-multiplexing tradeoff of MIMO coding, P_e is increasing on r. Therefore, we expect and will show rigorously later that there is a tradeoff between these two types of loss in the system. Our objective is to find the optimal multiplexing rate r^* that minimizes the total end-to-end loss probability.

IV. Problem Analysis

First, we obtain an analytical relation of the error probability due to MIMO channel, P_e, and the multiplexing rate r. By the diversity-multiplexing tradeoff for MIMO channel, we have that the probability of the whole T-symbol block in error is $\rho^{-d^*(r)}$. Since a timeslot contains exactly one block, when there is a decoding error, the data in the whole timeslot is in error. Hence, the average probability of bit loss due to MIMO channel is

$$P_e \doteq e^{-d^*(r) \log \rho}. \tag{11}$$

Since $d^*(r)$ is decreasing on the multiplexing rate r, it is clear that P_e is increasing on r.

Next, we obtain an analytical relation of the error probability due to delay bound violation, P_q, and the multiplexing rate r. Let Q be the steady-state queue length in the buffer. Since the queue is served at the constant rate C bits per timeslot, we define $B := DC$ be the queue length bound corresponding to the delay bound D, i.e.

$$B = rDT \log \rho. \tag{12}$$

Consider a bit who sees ahead of itself an amount of work greater than B. Such a bit is guaranteed to be delayed by more than D timeslots; hence, it will be obsolete. On the other hand, assuming a negligible propagation delay, such bits are the only bits lost due to delay violations. In other words, the probability of a bit becoming obsolete is nothing but the tail probability of the steady-state queue length, i.e. $P_q \approx \mathcal{P}(Q > B)$. Now we use the effective bandwidth result (discussed in Section II.B), as well as (5), (7), (8) and (12), to arrive at the following:

$$P_q \doteq e^{-\delta B} = e^{-\delta r D T \log \rho}, \tag{13}$$

where δ is the solution to

$$\alpha(\delta) = rT \log \rho = \frac{r\alpha(0)}{\lambda}. \tag{14}$$

Note that from (14) and the fact that $\alpha(s)$ is increasing in s, it is clear that δ is increasing in r; hence, P_q is decreasing in r. Now, by (11) and (13), we can rewrite the bound on the total loss probability as

$$\begin{aligned} P_t &\dot{\leq} e^{-\delta r DT \log \rho} + e^{-d^*(r) \log \rho} \\ &= e^{-\delta r DT \log \rho + o(\log \rho)} + e^{-d^*(r) \log \rho + o(\log \rho)} \end{aligned} \tag{15}$$

as $\log \rho \to \infty$. The terms in (15) provide us with an explicit characterization of the diversity-multiplexing tradeoff and its impact on the total loss probability, quite similar to [2]. The first term, corresponding to the delay bound violation, is decreasing in the multiplexing rate r, while the second term, corresponding to the channel error probability, is increasing with r. Hence, it is clear that there will be an optimal choice of $\lambda < r \leq \min(M, N)$.

V. Minimizing Total Loss Probability

A. Asymptotic Bound

To get analytical results for the optimal total loss probability bound, we consider the asymptotic case when $\rho \to \infty$. The minimum of P_t happens when the exponent of the two terms in (15) are within $o(1)$ of each other (note that if the exponents were not in the same order, one term would dominate in the

[3] For simplicity of presentation. The concept in this paper works for continuous system as well since the effective bandwidth holds there too.

sum as $\log \rho \to \infty$). In another word, the optimal multiplexing rate r^* happens when r^* is the root of

$$\alpha^{-1}(\frac{r\alpha(0)}{\lambda})rDT = d^*(r) + o(1), \tag{16}$$

where we substitute δ with $\alpha^{-1}(\frac{r\alpha(0)}{\lambda})$ by using (14). Notice that the optimal bound depends on the statistical characteristics of the source. This bound is illustrated in Figures 3 to 5 for various data sources discussed in Section VI.

B. Non-Asymptotic Bound

For practical systems, the SNR ρ is large but finite. If we assume that the asymptotic tail probability and decoding error probability hold for finite SNR, we can find the optimal diversity-multiplexing tradeoff by solving the following optimization problem:

$$\min_{\lambda < r \le \min(M,N)} \rho^{-\alpha^{-1}(\frac{r\alpha(0)}{\lambda})rDT} + \rho^{-d^*(r)} \tag{17}$$

The above optimizations are illustrated in Figures 3 to 5 for various data sources.

VI. Examples

A. I.I.D. Compound Poisson Sources

The arrival in a timeslot, X_t, is an aggregation of Poisson packet arrivals with general length distribution, i.e. $X_t = \sum_{n=1}^{N} Y_n$ where $Y_1, Y_2, \ldots$ are i.i.d. random variables with distribution G, and N is an independent Poisson random variable with rate ν packets per timeslot, then the effective bandwidth is [4]

$$\alpha(\delta) = \frac{\nu}{\delta}\int (e^{\delta x} - 1)dG(x). \tag{18}$$

The average bit arrivals per timeslot is ν/μ which is dictated by the source process, i.e. $\nu/\mu = \alpha(0)$.

Source I: Exponential-Length Poisson

Let $Y_1, Y_2, \ldots$ be exponentially distributed with mean$1/\mu$, then the effective bandwidth for this source is, by (18),

$$\alpha(\delta) = \begin{cases} \frac{\nu}{\mu-\delta} & \text{if } 0 \le \delta < \mu, \\ \infty & \text{if } \delta \ge \mu. \end{cases}$$

By using (14), and the fact that $\alpha(0) = \nu/\mu$, we obtain the error exponent δ of the tail probability as following:

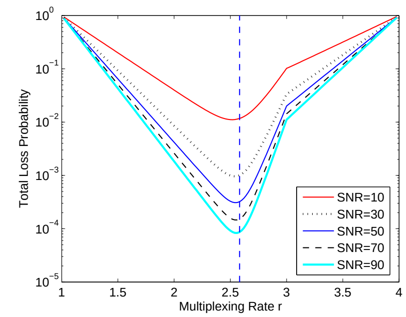
$$\delta = \mu(1 - \frac{\lambda}{r}). \tag{19}$$

Note that (19) confirms that the tail probability exponent δ is increasing in r.

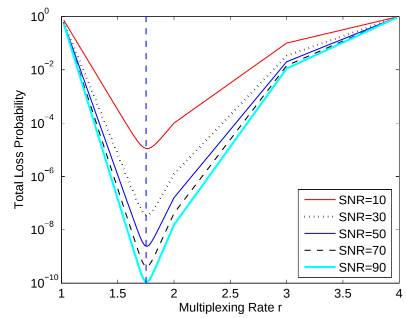
Source II: Fixed-Length Poisson

Let the packet length be deterministic and of size $1/\mu$, i.e. $Y_n = 1/\mu$, $\forall n$. From (18) the effective bandwidth for this source is

$$\alpha(\delta) = \frac{\nu(e^{\delta/\mu} - 1)}{\delta}.$$



(a) D=20



(b) D=100

Fig. 3. Source I: Total loss probability vs. multiplexing rate for $D = 20, 100$

B. Markov-Modulated Source

Source III: On-Off Exponential-Length Poisson

We consider a simple Markov modulated source: an on-off Markov source where $H_t \in \{\text{off}, \text{on}\}$. When H_t is on, arrivals $\{X_t\}$ are generated by a compound Poisson stream with exponential length at average rate ν/μ. When H_t is off, there are no arrivals. Suppose the transition probability matrix is $\begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix}$, where $0 < p, q < 1$, then the effective bandwidth for this on-off Poisson source is ([6] and [7])

$$\alpha(\delta) = \frac{1}{\delta}\log\left[\frac{1}{2}(a(\delta) + \sqrt{a^2(\delta) + 4b(\delta)})\right]$$

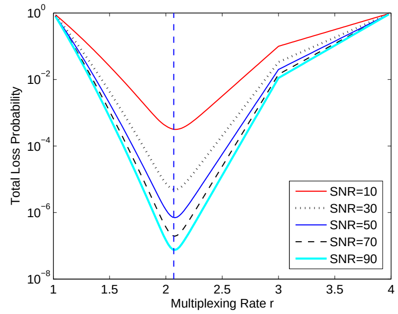
where

$$\begin{aligned} a(\delta) &= p + q\exp(\frac{\delta\nu}{\mu-\delta}), \\ b(\delta) &= (1-p-q)\exp(\frac{\delta\nu}{\mu-\delta}). \end{aligned}$$

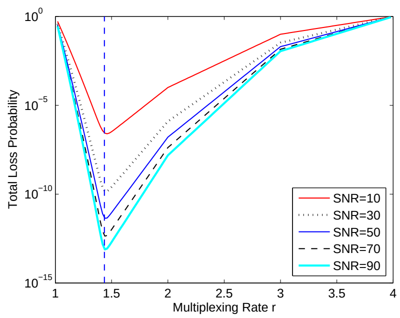
The average number of bit arrivals per timeslot is $(\frac{1-p}{2-p-q})\frac{\nu}{\mu}$.

C. Results

Figures 3 to 5 show the the total loss probability P_t vs. the multiplexing rate r for different levels of SNR ρ for Sources I-III, respectively. The vertical dashed lines show the optimal r^* obtained from the asymptotic case (16). For each source, we consider two delay bounds: $D = 20$ and 100. We assume $M = N = 4$ and $T = M + N - 1$. For all the sources, the average packet size $1/\mu = 100$ bits, and the average arrival rate $\alpha(0) = T\log\rho$, i.e. $\lambda = 1$. In Source III, we assume $p = q = 0.99$ so that the source is really bursty since the state

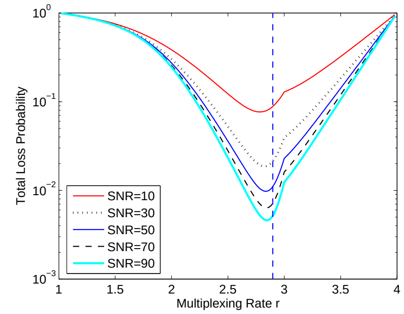


(a) D=20

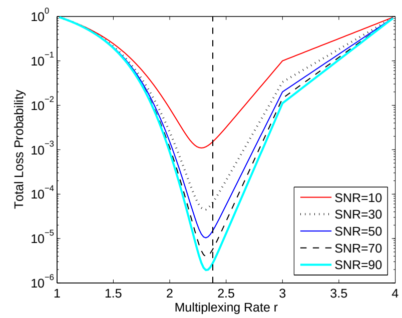


(b) D=100

Fig. 4. Source II: Total loss probability vs. multiplexing rate for $D = 20, 100$



(a) D=20



(b) D=100

Fig. 5. Source III: Total loss probability vs. multiplexing rate for $D = 20, 100$

transitions happen rarely. When Source III is in on state, it generates arrivals at the average rate twice of those in sources 1 and 2. If we rank the sources by their burstiness (standard deviation to average ratio), source II is the least bursty and III is the most. Note that in each figure, the optimal operating points are almost independent of the SNR values and very close to the r^* obtained from the asymptotic case.

The optimal operating multiplexing rates for the three sources are summarized in Figure 6. It shows that, given the same delay bound, the more bursty the source is, the higher the optimal multiplexing rate. Moreover, a less stringent delay bound D requires lower transmission rates. This figure summarizes the main intuitive relation between the source burstiness, the optimal operating multiplexing rate, and the delay bound.

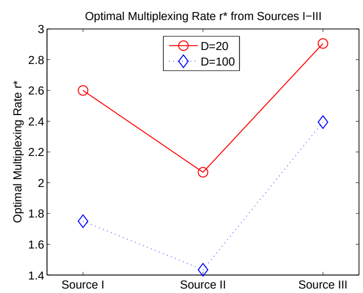


Fig. 6. Comparison of the optimal multiplexing rates r^* for Sources I-III.

VII. Conclusion and Future Studies

Based on the effective bandwidth representation of the source, we derive an analytical tradeoff between the end-to-end loss probability and the multiplexing rate and find the optimal multiplexing rate that minimizes this total loss probability. We demonstrate the technique by examples of i.i.d sources and a Markov-modulated source.

It would be of an interest to perform similar studies with varying queuing discipline and buffer size. The buffer size is of special interest since buffer overflow will introduce a third source of loss in the system.

References

[1] L. Zheng and D. Tse, "Diversity and multiplexing: a fundamental trade-off in multiple-antenna channels," *IEEE Trans. Info. Theory*, vol. 49, no. 5, May 2003.

[2] T. Holliday and A. Goldsmith, "Joint source and channel coding for MIMO systems," *Proc. Allerton Conf.*, 2004.

[3] R. Negi and S. Goel, "An information-theoretic approach to queuing in wireless channels with large delay bounds," *Globecomm'04*, 2004.

[4] F. Kelly, "Notes on effective bandwidths," *Stochastic Networks: Theory and Applications*, vol. 4, 1996.

[5] J. Walrand and P. Varaiya, *High-Performance Communication Networks*, 2nd ed. Morgan Kaufman, CA, 2000.

[6] D. Tse, R. G. Gallager, and J. N. Tsitsiklis, "Statistical multiplexing of multiple time-scale Markov streams," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, August 1995.

[7] G. Kesidis, J. Walrand, and C. S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," IEEE/ACM Trans. Networking, vol. 1, no. 4, pp. 424-428, Aug. 1993.

[8] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications*, Boston, MA: Jones and Bartlett, 1992.

[9] P. R. Jelenkovic, A. A. Lazar, and N. Semret, "The effect of multiple time scales and subexponentiality in MPEG video streams on queuing behavior," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 6, August 1997.